# COLUR: Confidence-Oriented Learning, Unlearning and Relearning with Noisy-Label Data for Model Restoration and Refinement

**Zhihao Sui**[1], **Liang Hu**[2*], **Jian Cao**[1*], **Usman Naseem**[3], **Zhongyuan Lai**[4], **Qi Zhang**[2]
[1]Shanghai Jiao Tong University
[2]Tongji University
[3]Macquarie University
[4]Shanghai Ballsnow Intelligent Technology Co. Ltd
fancyboy@sjtu.edu.cn, lianghu@tongji.edu.cn
cao-jian@sjtu.edu.cn, usman.naseem@mq.edu.au
abrikosoff@yahoo.com, zhangqi_cs@tongji.edu.cn

## Abstract

Large deep learning models have achieved significant success in various tasks. However, the performance of a model can significantly degrade if it is needed to train on datasets with noisy labels with misleading or ambiguous information. To date, there are limited investigations on how to restore performance when model degradation has been incurred by noisy label data. Inspired by the "forgetting mechanism" in neuroscience, which enables accelerating the relearning of correct knowledge by unlearning the wrong knowledge, we propose a robust model restoration and refinement (MRR) framework COLUR, namely Confidence-Oriented Learning, Unlearning and Relearning. Specifically, we implement COLUR with an efficient co-training architecture to unlearn the influence of label noise, and then refine model confidence on each label for relearning. Extensive experiments are conducted on four real datasets and all evaluation results show that COLUR consistently outperforms other SOTA methods after MRR.

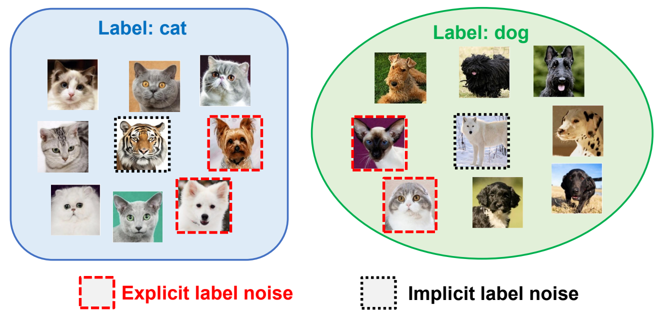


Figure 1: An illustration of labeling on two classes: cat and dog. The label noise mainly comes from two causes: mislabeling and ambiguous labeling on images.

## 1 Introduction

AI models have achieved remarkable success in computer vision tasks due to the availability of massive datasets and advanced computational resources. In general, the performance of a trained model is heavily dependent on the quality of labeling on training data. The models need to keep learning new knowledge from the data generated every day. However, massive label noise inevitably occurs with the new data. As a result, model updating on the new data with heavy label noise will inevitably lead to performance degradation.

According to the analysis above, label noise is a major factor contributing to the degradation of model performance. As shown in Figure 1, the label noise in a dataset mainly comes from two sources: **(a)** mislabeled data from human or automatic annotators, namely explicit label noise; **(b)** hard labels on ambiguous data, namely implicit label noise. Obviously, it is not practical to manually correct and refine the labels from a huge amount of training data. Moreover, fully retraining a degraded task model from scratch is also time-consuming and computationally expensive. To mitigate the influence of label noise, especially explicit label noise, Learning with Noisy Labels (LNL) [Algan and Ulusoy, 2021], is a highly relevant research area that has attracted more attention in recent years. The goal of LNL is mainly to train the model from scratch on the basis of data with noisy labels. As stated above, since retraining a model from scratch is generally impossible in real-world situations, we aim to *restore and refine a trained model that has suffered degradation*.

In general, LNL methods can be grouped into two categories. One is the model-based approach, in which various LNL model architectures are constructed for label correction [Li *et al.*, 2021; Ortego *et al.*, 2021; Zhang *et al.*, 2021b], and sample selection [Li *et al.*, 2020; Cheng *et al.*, 2021; Karim *et al.*, 2022]. However, this approach is not suitable for real AI systems because different task models may have different architectures and have been fine-tuned for specific tasks. As a result, the essential requirement is to refine the performance of the degraded task models instead of replacing them with LNL models. The second category is the model-free approach to improve the robustness of the model against

*Corresponding author

label noise by robust losses [Liu *et al.*, 2020; Xu *et al.*, 2019; Wang *et al.*, 2019] or regularizers [Englesson and Azizpour, 2021; Sarfraz *et al.*, 2021]. However, this approach is typically designed for the model learning process over noisy label data, which is invalidly used for the restoration of the degraded task models. Although existing LNL approaches have limitations in *model restoration and refinement* (**MRR**), the underlying building blocks can be easily borrowed to construct the proposed MRR framework.

To efficiently implement MRR, we need to mitigate the influence on a model w.r.t. those noisy label data and relearn the data with refined labels. Inspired by neuroscience research [Gong *et al.*, 2024; Bao *et al.*, 2025], *Forgetting* has been experimentally demonstrated for efficient relearning after learning [Ryan and Frankland, 2022]. We argue that Machine Unlearning (MU) [Xu *et al.*, 2023] is a machine forgetting [Sha *et al.*, 2024] mechanism that provides a possible way to unlearn the influence of label noise on degraded models, avoiding retraining the model from scratch. After unlearning, it is easier and more effective to relearn the model using LNL techniques. In fact, the "Learning, Unlearning and Relearning (**LUR**)" theory has demonstrated its success in various areas, including education [Klein, 2008], self-training [Dunlap and Lowenthal, 2011], and online learning for deep neural networks [Ramkumar *et al.*, 2023; Miao *et al.*, 2024]. As a result, we propose a **Confidence-Oriented LUR (COLUR)** framework to efficiently and effectively address the challenges of MRR over noisy label data.

Specifically, the proposed COLUR is a model-agnostic MRR framework, which aims to refine the confidence over noisy label data samples for training AI models with different architectures. Inspired by the co-training architecture [Han *et al.*, 2018; Yu *et al.*, 2019; Jiang *et al.*, 2018] in LNL, we additionally duplicate the task model architecture and initiate a teacher model. Given any degraded model learned on noisy label data, we first perform machine unlearning to reduce the model confidence on those data with high disagreement scores measured by teacher and student models. Then we relearn the unlearned model over an augmented dataset by mixing low-confidence label data with high-confidence label data [Carratino *et al.*, 2022; Zhang *et al.*, 2018]. Through the unlearning and relearning process, COLUR can effectively restore and refine model performance. The contributions of this work are summarized as follows:

- We insightfully study a critical but inadequately investigated problem of model restoration and refinement (MRR) on widespread noisy label datasets.
- We propose a robust MRR framework, COLUR, in view of the "learning, unlearning and relearning" mechanism that benefits relearning from misinformation.
- We implement COLUR with a co-training architecture to iteratively refine model confidences to unlearn the impact from noisy labels and select confident labels for relearning.
- Extensive experiments are conducted on a collection of real datasets to demonstrate the superior performance of COLUR in restoring the degraded models and to visualize the results of label correction.

## 2 Related Work

LNL and MU methods are two key references for COLUR.

### 2.1 Learning with Noisy Labels

LNL mainly aims to train a model based on data with noisy labels [Algan and Ulusoy, 2021], which can be divided into two types of approaches. One is the model-free approach to improve the robustness of the model against label noise by robust losses [Ghosh *et al.*, 2017; Xu *et al.*, 2019; Zhang and Sabuncu, 2018; Zhang *et al.*, 2021a], such as GJS [Englesson and Azizpour, 2021] using the generalized Jenson-Shannon divergence as a loss, or regularization methods [Zhou *et al.*, 2021; Sarfraz *et al.*, 2021], such as ELR [Liu *et al.*, 2020] adopting the early-learning regularization strategy. The other is the model-based approach for label correction and sample selection. Co-training is one of the representative methods to conduct label refinement in terms of agreements and/or disagreements between two models. In this way, Co-teaching [Han *et al.*, 2018] and Co-teaching+[Yu *et al.*, 2019] train two DNNs, where Co-teaching selects training samples with agreement, and Co-teaching+ selects samples with disagreement based on the idea of Decoupling [Malach and Shalev-Shwartz, 2017]. DivideMix [Li *et al.*, 2020] uses Gaussian mix models to partition the original dataset, employing the semi-supervised technique MixMatch to generate mixed training data. JoCoR [Wei *et al.*, 2020] extends Co-teaching with a data augment strategy to align teacher and student models with agreements. Label correction focuses on correcting the labels of noisy samples. NLNL [Kim *et al.*, 2019] employs negative learning to refine the classification boundary of noisy label data. DISC [Li *et al.*, 2023] proposes a dynamic threshold strategy that focuses on the memory strength of DNN to select and correct for noisy labels.

In view of the inefficiency of LNL for MRR, we propose COLUR to address the challenges of MRR in terms of the "learning, unlearning and relearning" mechanism.

### 2.2 Machine Unlearning

Machine Unlearning (MU) refers to the process of selectively removing the influence of specific training data from a trained model [Warnecke *et al.*, 2021; Graves *et al.*, 2021; Golatkar *et al.*, 2020; Becker and Liebig, 2022; Izzo *et al.*, 2021]. Model fine-tuning (FT) [Warnecke *et al.*, 2021] achieves unlearning by fine-tuning the retained dataset, but may lead to catastrophic forgetting [Kirkpatrick *et al.*, 2017]. Gradient ascent (GA) [Graves *et al.*, 2021; Golatkar *et al.*, 2020] reverses the model training process by moving the model parameters in the direction where the loss of erased data increases. Typical methods can remove the influence of the specified sample on the model parameters by modifying the fisher information matrix [Becker and Liebig, 2022] or the influence function (IU) [Koh and Liang, 2017; Izzo *et al.*, 2021]. In addition, the $\ell 1$-sparse based weight pruning (L1-WP) adopted to improve the sparsity of the model could improve the effectiveness of data erasure [Liu *et al.*, 2024]. Some recent work has explored more precise unlearning on target forget instances. Unlearning via Null Space Calibration (UNSC) [Chen *et al.*, 2024] constrains the

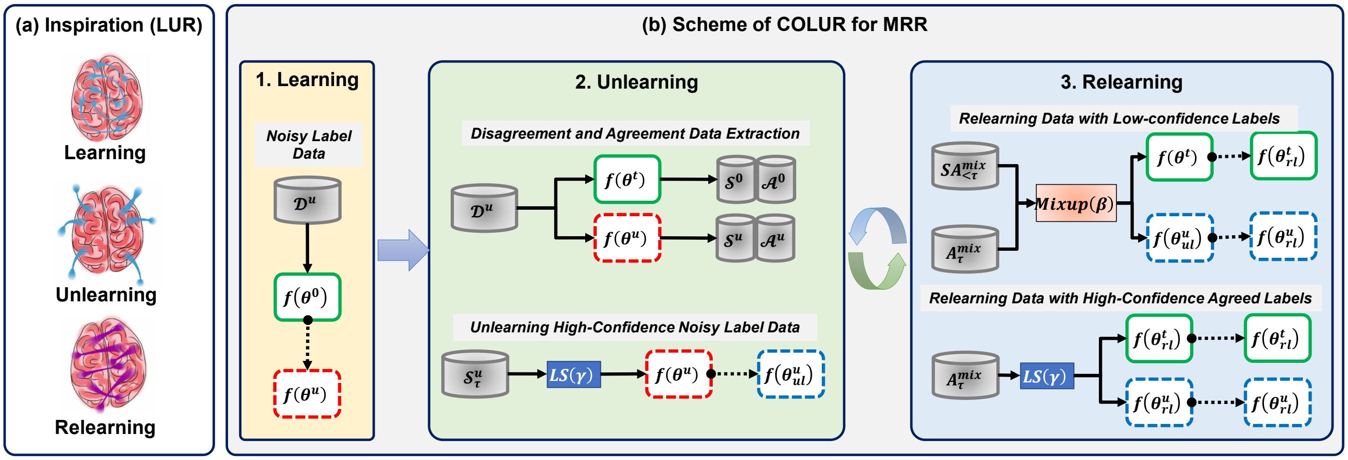


Figure 2: **(a) Inspiration of LUR and (b) Scheme of COLUR Scheme for MRR**. (1) **Learning:** The model $f(\theta^0)$ is incrementally trained on a noisy label dataset leading to a degraded model $f(\theta^u)$; (2) **Unlearning:** MU is employed to unlearn high-confidence noisy label data, resulting in model $f(\theta^u_{ul})$; (3) **Relearning:** the confidence-refined dataset is constructed for relearning $f(\theta^u_{ul})$, finally leading to the refined model $f(\theta^u_{rl})$. The iterative process of Unlearning and Relearning is executed alternatively.

unlearning process within a null space tailored to the remaining samples to ensure that unlearning does not negatively impact the performance of the model. SalUn [Fan *et al.*, 2024] introduces the concept of "weight saliency" to narrow the performance gap with exact unlearning.

Although MU methods can be used as a tool to mitigate the influence of label noise for MRR, they are limited in practice because they require prior clarification of which data points need to be deleted. In addition, unlearning processes can unintentionally degrade model performance and can lead to "*over-unlearning*" [Chen *et al.*, 2024].

# 3 Proposed Method

## 3.1 Preliminaries

### Problem Formulation

Consider a training dataset $\mathcal{D}^0 = \{(x_i^0, y_i^0)\}_{i=1}^N$, where $x_i^0 \in \mathcal{X}^0$ are input samples and $y_i^0 \in \mathcal{Y}^0$ are the corresponding labels. A task model $f(\theta^0)$ is trained by minimizing the following classification loss:

$$\theta^0 = \arg\min_\theta \mathcal{L}(f(\mathcal{X}; \theta), \mathcal{Y}). \tag{1}$$

Now, let us denote a new training dataset with noisy labels $\mathcal{D}^u = \{(x_i^u, y_i^u)\}_{i=1}^M$ where each $y_i^u \in \mathcal{Y}^u$ may be an incorrect label on $x_i$. Then, a degraded model $f(\theta^u)$ can be obtained by incrementally trained on $\mathcal{D}^u$:

$$\theta^u = \arg\min_\theta \mathcal{L}(f(\mathcal{X}^u; \theta|\theta^0), \mathcal{Y}^u). \tag{2}$$

The goal of this paper is to find an optimal $\theta^u_{rl}$ that can restore and refine the performance of the degraded model $f(\theta^u)$ based on $\mathcal{D}^u$:

$$\theta^u_{rl} = \arg\min_\theta \mathcal{L}(f(\mathcal{D}^u; \theta|\theta^u)). \tag{3}$$

### Definitions

Model confidence [Kim *et al.*, 2019; Li *et al.*, 2023] and label smoothing [Szegedy *et al.*, 2016] play an important role in later sections, so we define them first.

**Definition 1.** *(Model Confidence): Given a model $f(\theta)$, and a sample $x$, the confidence of the model $c(\theta)$ w.r.t. $x$ is defined as:*

$$c(x; \theta) = \max(\mathbf{p}), \ \textit{where}\ \mathbf{p} = f(x; \theta) \tag{4}$$

$\mathbf{p}$ *is the probabilities over all classes output by $f(\theta)$.*

**Definition 2.** *(Label Smoothing): Given a hard label $y$, the label smoothing function is defined as:*

$$\tilde{\mathbf{y}}^\gamma = LS(y; \gamma) = (1-\gamma)\cdot \mathbf{y} + \frac{\gamma}{K}\cdot \mathbf{1} \tag{5}$$

*where $\mathbf{y}$ denotes the one-hot vector of $y$, $\mathbf{1}$ stands for an all-one vector, $K$ is the number of label classes, and $\gamma \in [0, 1]$ is the smooth rate.*

## 3.2 Framework Overview

Figure 2 shows the workflow of COLUR. First, given a task model $f(\theta^0)$ that was originally trained on a dataset $\mathcal{D}^0$ (cf. Eq (1)), the incremental "**learning**" process is conducted on a real-world dataset with label noise, and leading to $f(\theta^u)$. If significant performance degradation (e.g. 10%) occurs in $f(\theta^u)$ (cf. Eq (2)), we start the "**unlearning and relearning**" process to restore the performance of the degraded model, and finally obtain the refined task model $f(\theta^u_{rl})$ (cf. Eq (3)). In particular, we adopt a co-training strategy to conduct the unlearning and relearning process, where $f(\theta^t)$ serves as the teacher model and $f(\theta^u)$ serves as the student model.

## 3.3 Details of COLUR for MRR

In this section, we present the details of the "Learning, Unlearning and Relearning (LUR)" scheme for MRR.

### (1) Learning on Noisy Label Data

We can obtain the performance-degraded model $f(\theta^u)$ if the task model $f(\theta^0)$ is incrementally trained on the dataset $\mathcal{D}^u$ (cf. Eq (2)), i.e. learning on noisy label data as shown in Figure 2 (b).

**(2) Unlearning Impact of Noisy Label Data**

According to the LUR theory [Klein, 2008; Dunlap and Lowenthal, 2011; Ramkumar *et al.*, 2023], **unlearning** aims to eliminate highly uncertain knowledge and assumptions to allow new insights, while relearning builds on previous experiences. Therefore, we need to identify the most likely mislabeled data and perform unlearning to mitigate the influence on the degraded model $f(\theta^u)$ and then relearning on the label-refined data.

**Data with High-Confidence Disagreed Labels.** To identify the data associated with high-confidence label noise, we first collect predictions (labels and probabilities) from both the teacher model $f(\theta^t)$ (note that $\theta^t$ is initialized with the copy of $\theta^0$) and the student model $f(\theta^u)$ with input $\mathcal{X}^u$:

$$\mathcal{Y}^t, \mathcal{P}^t = f(\mathcal{X}^u; \theta^t), \quad \mathcal{Y}^u, \mathcal{P}^u = f(\mathcal{X}^u; \theta^u) \tag{6}$$

Then, we extract the disagreement sets according to predictive labels from the teacher model and the student model:

$$\begin{aligned} \mathcal{S}^t = \{(x, y^t, \mathbf{p}^t)\},\ \mathcal{S}^u = \{(x, y^u, \mathbf{p}^u)\} \\ \text{for}\ \ y^t(x) \neq y^u(x); x \in \mathcal{X}^u \end{aligned} \tag{7}$$

If the confidence scores $c(x;\theta^t)$ and $c(x;\theta^u)$ (cf. Eq (4)) of both models are higher, the disagreement on $x$ tends to have a higher confidence. Accordingly, we define the joint confidence score in terms of their geometric mean:

$$c(x; \theta^t, \theta^u) = \sqrt{c(x;\theta^t) \cdot c(x;\theta^u)} \tag{8}$$

Intuitively, the data samples with higher confidence disagreement indicate higher uncertainty, i.e., higher probability of noise. As a result, we can obtain the high-confidence noisy label datasets, $\mathcal{S}^u_\tau, \mathcal{S}^t_\tau$, at level $\tau$:

$$\mathcal{S}^u_\tau = \{(x, y^u, \mathbf{p}^u) | c(x; \theta^t, \theta^u) \geq \tau\} \tag{9}$$

$$\mathcal{S}^t_\tau = \{(x, y^t, \mathbf{p}^t) | c(x; \theta^t, \theta^u) \geq \tau\} \tag{10}$$

**Unlearning High-Confidence Noisy Label Data.** According to the LUR theory, we need to mitigate the influence of the above high-confidence noisy label data, $\mathcal{S}^u_\tau$, from the degraded model $f(\theta^u)$. Label smoothing (LS) is an effective tool for dealing with overconfidence and label noise [Szegedy *et al.*, 2016; Lukasik *et al.*, 2020; Wei *et al.*, 2022]. Recent work [Di *et al.*, 2024] has shown that LS can improve MU by making wrong predictions with equally low confidence. As a result, we adopt an LS-based gradient ascent (GA) [Di *et al.*, 2024] to unlearn the influence of $\mathcal{S}^u_\tau$.

$$(\mathcal{X}_\tau, \tilde{\mathcal{Y}}^\gamma_\tau) : \{(x, \tilde{\mathbf{y}}^\gamma) | \tilde{\mathbf{y}}^\gamma = LS(y^u; \gamma)\} \tag{11}$$

$$\theta^u_{ul} = \theta^u_{ul} + \lambda_u \frac{\partial \mathcal{L}(f(\mathcal{X}_\tau; \theta^u_{ul}), \tilde{\mathcal{Y}}^\gamma_\tau)}{\partial \theta^u_{ul}} \tag{12}$$

**(3) Relearning with Label Confidence Refinement**

Relearning after unlearning is an efficient way to rebuild knowledge systems [Ryan and Frankland, 2022; Sha *et al.*, 2024]. Consequently, it is necessary to relearn the new knowledge from the noisy label dataset $\mathcal{D}^u$ with label refinement based on the above unlearned model $f(\theta^u_{ul})$ (cf. Eq (12)).

**Data with High-Confidence Agreed Labels.** The agreement datasets can be obtained as Eq (7):

$$\begin{aligned} \mathcal{A}^t : \{(x, y, \mathbf{p}^t)\},\ \mathcal{A}^u = \{(x, y, \mathbf{p}^u)\} \\ \text{for}\ \ y = y^t(x) = y^u(x); x \in \mathcal{X}^u \end{aligned} \tag{13}$$

The joint agreement model confidence score on input $x$ is computed by:

$$c(x; \theta^u_{ul}, \theta^t) = \sqrt{c(x;\theta^u_{ul}) \cdot c(x;\theta^t)} \tag{14}$$

Similarly to Eqs (9) and (10), we obtain the high confidence agreement datasets $\mathcal{A}^t_\tau$ and $\mathcal{A}^u_\tau$ where $c(x;\theta^u_{ul},\theta^t) > \tau$. Furthermore, we define $\mathcal{A}^{mix}_\tau : \{(x, \mathbf{p}^{mix}_\tau)\}$ by the average probability of $\mathcal{A}^t_\tau$ and $\mathcal{A}^u_\tau$, i.e., $\mathbf{p}^{mix}_\tau = (\mathbf{p}^t + \mathbf{p}^u)/2$.

**Relearning Data with Low-Confidence Labels.** Let us denote the union of the low-confidence disagreement dataset and the low-confidence agreement dataset w.r.t. teacher model and student model as: $\mathcal{SA}^t_{<\tau} = \mathcal{S}^t_{<\tau} \cup \mathcal{A}^t_{<\tau}$and $\mathcal{SA}^u_{<\tau} = \mathcal{S}^u_{<\tau} \cup \mathcal{A}^u_{<\tau}$. For each sample $x$, we mix the low confidence predictive probabilities of $\mathcal{SA}^t_{<\tau}$ and $\mathcal{SA}^u_{<\tau}$, and obtain $\mathcal{SA}^{mix}_{<\tau} : \{(x, \mathbf{p}^{mix}_{<\tau})\}$:

$$\mathbf{p}^{mix}_{<\tau} = \beta_m \cdot \mathbf{p}^t_{<\tau} + (1 - \beta_m) \cdot \mathbf{p}^u_{<\tau} \tag{15}$$

where $\beta_m \sim Beta(\alpha_m, \alpha_m)$. Recent work has shown that Mixup [Zhang *et al.*, 2018] is an effective regularization technique to deal with label noise [Carratino *et al.*, 2022] defined as follows, where $\beta \sim Beta(\alpha, \alpha)$.

$$\begin{aligned} (\tilde{x}, \tilde{\mathbf{p}}) = Mixup\big((x_1, \mathbf{p}_1), (x_2, \mathbf{p}_2)\big) \\ \tilde{x} = \beta x_1 + (1-\beta)x_2,\ \ \tilde{\mathbf{p}} = \beta \mathbf{p}_1 + (1-\beta)\mathbf{p}_2 \end{aligned} \tag{16}$$

To refine the data with low-confidence labels, we mix them with the data with high-confidence labels.

$$(\mathcal{X}^{mix}, \mathcal{P}^{mix}) = Mixup(\mathcal{SA}^{mix}_{<\tau}, \mathcal{A}^{mix}_\tau) \tag{17}$$

Then, we relearn teacher and student models with the above mixup data based on the unlearned model $f(\theta^u_{ul})$:

$$\theta^u_{rl} = \theta^u_{rl} - \lambda_u \frac{\partial \mathcal{L}(f(\mathcal{X}^{mix}; \theta^u_{rl} | \theta^u_{ul}), \mathcal{P}^{mix})}{\partial \theta^u_{rl}} \tag{18}$$

$$\theta^t_{rl} = \theta^t_{rl} - \lambda_t \frac{\partial \mathcal{L}(f(\mathcal{X}^{mix}; \theta^t_{rl} | \theta^t), \mathcal{P}^{mix})}{\partial \theta^t_{rl}} \tag{19}$$

**Relearning Data with High-Confidence Agreed Labels.** In general, the data with high-confidence agreed labels, i.e. $\mathcal{A}^t_\tau$ and $\mathcal{A}^u_\tau$, imply correct labels with high probability. As a result, they can serve as high-quality representative samples for relearning. To avoid over-confidence in these data, we relearn $\mathcal{A}^{mix}_\tau : \{\mathcal{X}_\tau, \mathcal{Y}_\tau\}$ with label smoothing:

$$\theta^u_{rl} = \theta^u_{rl} - \lambda_u \frac{\partial \mathcal{L}(f(\mathcal{X}_\tau; \theta^u_{rl}), LS(\mathcal{Y}_\tau; \alpha))}{\partial \theta^u_{rl}} \tag{20}$$

$$\theta^t_{rl} = \theta^t_{rl} - \lambda_t \frac{\partial \mathcal{L}(f(\mathcal{X}_\tau; \theta^t_{rl}), LS(\mathcal{Y}_\tau; \alpha))}{\partial \theta^t_{rl}} \tag{21}$$

As shown in Figure 2, the steps **(2) Unlearning** and **(3) Relearning** are run alternately. More details about the algorithm of COLUR can be found in the online extended version.

| Methods | CIFAR-10 (sym) | | | | CIFAR-100 (asym) | | | | Flower-102 (sym) | | | | Oxford-IIIT Pet (asym) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10% | 25% | 75% | 90% | 10% | 25% | 75% | 90% | 10% | 25% | 75% | 90% | 10% | 25% | 75% | 90% |
| Degrade | 84.82 | 79.30 | 45.82 | 30.04 | 63.70 | 55.88 | 17.37 | 7.50 | 93.63 | 87.35 | 10.69 | 2.84 | 90.73 | 82.72 | 18.89 | 4.93 |
| CoTe. | 45.12 | 44.70 | 46.28 | 53.70 | 37.75 | 40.93 | 40.45 | 38.57 | 90.88 | 64.41 | 16.96 | 3.53 | 86.21 | 67.62 | 66.15 | 50.83 |
| CoTe.+ | 73.29 | 73.39 | 77.44 | 70.64 | 62.12 | 58.36 | 52.78 | 51.39 | 89.31 | 85.29 | 14.02 | 2.65 | 88.91 | 88.99 | 69.96 | 28.62 |
| Decoup. | 83.52 | 82.44 | 77.66 | 71.99 | 60.27 | 59.73 | 56.75 | 51.28 | 92.55 | 86.27 | 15.69 | 3.63 | 89.18 | 88.44 | 79.45 | 47.86 |
| DISC | 85.37 | 85.15 | 79.64 | 76.39 | 61.48 | 60.73 | 57.02 | 56.29 | 91.27 | 87.35 | 62.90 | 29.41 | 73.24 | 67.73 | 54.97 | 35.92 |
| ELR | 82.63 | 81.96 | 78.06 | 73.74 | 59.04 | 59.23 | 54.72 | 55.11 | 90.39 | 87.75 | 40.49 | 15.59 | 76.83 | 72.25 | 60.70 | 35.32 |
| GJS | 85.13 | 84.10 | 79.58 | 75.17 | 61.15 | 60.97 | 56.89 | 54.47 | 90.10 | 89.90 | 58.33 | 22.16 | 73.83 | 72.99 | 54.54 | 33.42 |
| JoCoR | 85.75 | 83.31 | 64.59 | 54.23 | 61.58 | 50.63 | 36.44 | 38.54 | 92.06 | 85.20 | 13.63 | 2.75 | 49.93 | 84.68 | 42.90 | 28.40 |
| NLNL | 84.28 | 83.80 | 79.84 | 75.40 | 61.40 | 61.37 | 55.99 | 54.61 | 93.63 | 90.69 | 16.76 | 3.92 | 91.25 | 90.43 | 85.58 | 63.59 |
| PENCIL | 86.21 | 84.47 | 74.62 | 64.62 | 65.02 | 61.49 | 39.84 | 17.64 | 93.82 | 87.45 | 10.98 | 2.75 | 92.01 | 85.20 | 15.18 | 3.62 |
| COLUR | **88.53** | **87.74** | **84.12** | **80.34** | **66.78** | **65.43** | **60.95** | **58.26** | **94.61** | **90.39** | **80.20** | **58.10** | **92.59** | **92.26** | **89.18** | **78.39** |

Table 1: **Performance Comparison of MRR under Different Noise Levels**. The noise ratios of at 25% and 75% (normal case), and 10% and 90% (extreme case) correspond to the percentages of $|D_n^u| : |D^u|$. **Degrade** represents the performance after training on the noisy label datasets $\mathcal{D}^u$ with varying noise levels. The second section includes SOTA LNL methods for comparison, and the final row shows the performance of our method, **COLUR**. The best result from each group is highlighted in **bold**, while the second-best one is underlined.

| Dataset | $\|\mathcal{D}^{tr}\|$ | $\|\mathcal{D}^{ts}\|$ | $\|\mathcal{D}^0\| : \|\mathcal{D}^u\|$ | Noise |
|---|---|---|---|---|
| CIFAR-10 | 50,000 | 10,000 | 40% : 60% | sym. |
| CIFAR-100 | 50,000 | 10,000 | 60% : 40% | asym. |
| Flower-102 | 6,049 | 1,020 | 40% : 60% | sym. |
| Oxford-IIIT Pet | 3,680 | 3,669 | 30% : 70% | asym. |

Table 2: **Data Preparation.** For noise types, 'sym.' and 'asym.' represent symmetric and asymmetric noise, respectively.

# 4 Experiment

## 4.1 Experiment Setups

**Dataset Preparation.** We use four datasets in this experiment: **CIFAR-10 (C-10)** [Krizhevsky *et al.*, 2009], **CIFAR-100 (C-100)** [Krizhevsky *et al.*, 2009], **Flower-102 (F-102)** [Nilsback and Zisserman, 2008], and **Oxford-IIIT Pet (P-37)** [Parkhi *et al.*, 2012]. CIFAR-10 and CIFAR-100 consist of 60,000 low-resolution images classified into 10 and 100 classes, respectively. Flower-102 has 8,189 high-resolution images in 102 classes, and Oxford-IIIT Pet contains 7,349 high-resolution images of cats and dogs in 37 classes. These datasets cover both low- and high-resolution data, providing a comprehensive evaluation.

As shown in Table 2, we use the official split of the train set $\mathcal{D}^{tr}$ and test set $\mathcal{D}^{ts}$ in the dataset package. Each train set $\mathcal{D}^{tr}$ is further divided into initial training data $\mathcal{D}^0$ and incremental training data $\mathcal{D}^u$. $\mathcal{D}^u$ is further divided into a noisy label subset $\mathcal{D}_n^u$ and a clean label subset $\mathcal{D}_c^u$, with noise ratios. For noise types, **symmetric noise** is applied to CIFAR-10 and Flower-102 by randomly changing labels to any other class. In CIFAR-100 and Oxford-IIIT Pet, we apply **asymmetric noise** by changing labels to another within the same superclass, simulating real-world mislabeling. This setup ensures both *diversity* and *balance* in noise types in datasets. More details can be found in the online extended version.

**Comparison Methods.** To comprehensively compare the performance of MRR, a set of representative SOTA LNL methods are involved, including **(1)** Robust Loss Approach: **GJS** [Englesson and Azizpour, 2021]; **(2)** Regularization Approach: **ELR** [Liu *et al.*, 2020]; **(3)** Co-training Approach: Decoupling (**Decoup.**) [Malach and Shalev-Shwartz, 2017], Co-teaching (**CoTe.**) [Han *et al.*, 2018], Co-teaching+(**CoTe.+**) [Yu *et al.*, 2019], and **JoCoR** [Wei *et al.*, 2020]; **(4)** Label Correction Approach: **PENCIL** [Yi and Wu, 2019], **NLNL** [Kim *et al.*, 2019] and **DISC** [Li *et al.*, 2023];.

**Training Details.** Since the proposed COLUR framework is model-agnostic, we use EfficientNet [Tan and Le, 2019] and WideResNet [Zagoruyko and Komodakis, 2016] as the backbone for different datasets, as detailed in Table 2. To better train different comparison models, we try the AdamW and SGD optimizers with a momentum of 0.9 and a weight decay of 0.001. Other more detailed settings can be found in the online extended version. For each comparison model, we carefully tuned its hyperparameters.

## 4.2 Comparison with SOTA Methods

### Quantitative Results

To comprehensively compare the robustness between the proposed COLUR and different types of SOTA LNL methods as presented in the last subsection, we respectively set the noise ratios of labels at 25% and 75% (normal case), and 10% and 90% (extreme case). In addition, we also reported the performance of **Degrade** models after training on $\mathcal{D}^u$ as the reference to illustrate the capability of model restoration. According to the results reported in Table 1, we observe that the performance of all comparison methods generally decreases as the noise ratio increases. As expected, a higher level of label noise leads to higher uncertainty, increasing the difficulty of model classification. In particular, our COLUR method consistently achieves better performance across all noise levels compared to other comparison methods.

For example, COLUR achieves the best MRR accuracy of 87.74% (CIFAR-10) and 92.26% (Oxford-IIIT Pet), outperforming other baselines such as the second-best methods, DISC 85.15% (CIFAR-100) and NLNL 90.43% (Oxford-IIIT Pet) at a noise ratio of 75%. On other datasets with different levels, COLUR also demonstrates the superiority of our method in handling various levels of label noise. We can find that all the comparison LNL methods tend to perform better at lower noise ratios but struggle as the noise increases,

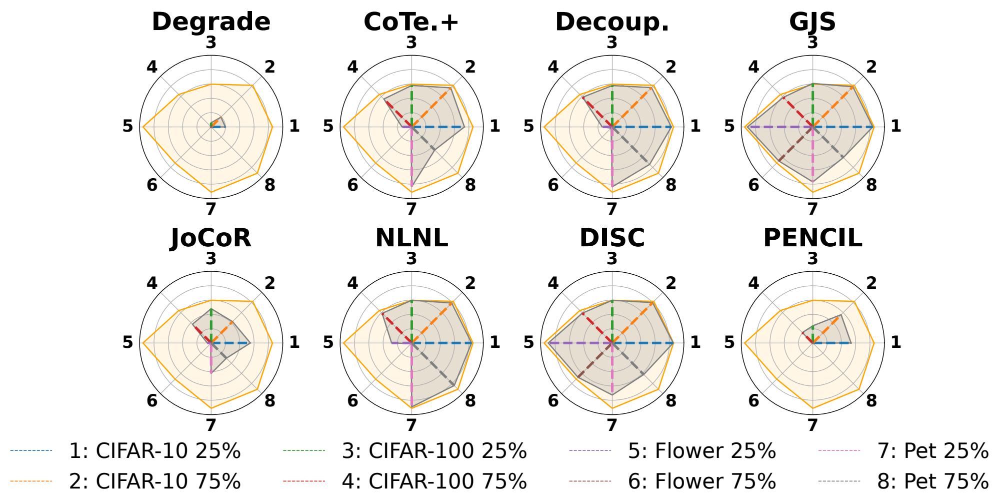


Figure 3: **The Accuracies After Label Correction**. Each radar chart compares the accuracies after the model restoration on 25% and 75% noise-ratio datasets: CIFAR-10, CIFAR-100, Flower-102, and Oxford-IIIT Pet. The Gray-shaded area represents each method's performance, with a larger area indicating better noise-handling capability. Our **COLUR** method, outlined in Orange, consistently covers a broader area compared to LNL and MU methods, demonstrating its robustness across datasets and noise levels.

especially in the case of asymmetric noise, where label corruption is more systematic. Overall, COLUR has experimentally demonstrated higher robustness and effectiveness across different noise levels, further confirming its applicability in real-world scenarios with highly noisy label data.

### Visualization of Label Correction

Figure 3 visualizes the accuracy after MRR for comparison methods, where each axis represents a different noise ratio (25% and 75%) with respect to each dataset. From the radar charts, it is evident that the proposed COLUR method consistently maintains a larger area compared to other baseline methods, indicating the superiority of label noise correction for all datasets and noise levels. As presented above, the comparison LNL methods show decent performance at lower noise levels, but their effectiveness diminishes as the noise ratio increases. These visualizations further confirm that the proposed COLUR is more robust and effective across various datasets at different levels of label noise, highlighting its advantage in label correction in terms of the LUR mechanism.

### Further Analysis on Extreme Cases

To assess the robustness of our methods and other baselines, we further analyze all comparison methods under extreme label noise conditions: one is the case of very low label noise with a ratio of 10% and the other is the case of very high label noise with a ratio of 90%. The results of four datasets have been reported in Table 1.

Specifically, in the case of a very low label noise ratio (10%), most comparison methods perform reasonably well, as the majority of the data are clean. However, our method still achieves the highest accuracy, which indicates that COLUR can not only correct explicit mislabels, but also refine implicit label noise on ambiguous images. In the very high label noise ratio (90%), the performance of most LNL methods deteriorates significantly due to the overwhelming presence of noisy labels. In particular, COLUR still maintains better performance compared to other methods, demonstrating its robustness in handling extremely noisy label datasets. This suggests that our unlearning and relearning strategy effectively mitigates the impact of severe label noise.

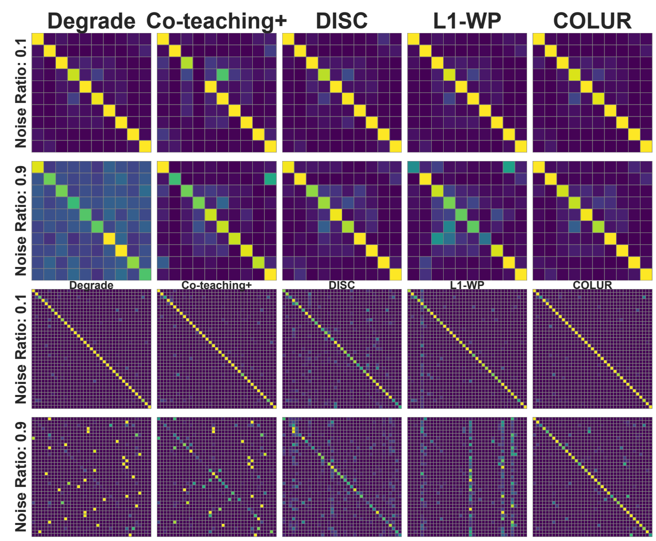


Figure 4: **Confusion Matrices under Very Low and High Noise Levels**. The top two rows display the confusion matrices for CIFAR-10, while the bottom two rows for Oxford-IIIT Pet. Each matrix compares the performance of various methods at low (10%) and high (90%) label noise ratios. Diagonal elements represent correct classifications, while off-diagonal elements indicate mislabeling.

**Visualization of Confusion Matrices.** The confusion matrices in Figure 4 show the classification performance in noisy label sets $\mathcal{D}_n^u$ for different methods at different noise ratios (10% and 90%). At a very low noise level (10%), most meth-

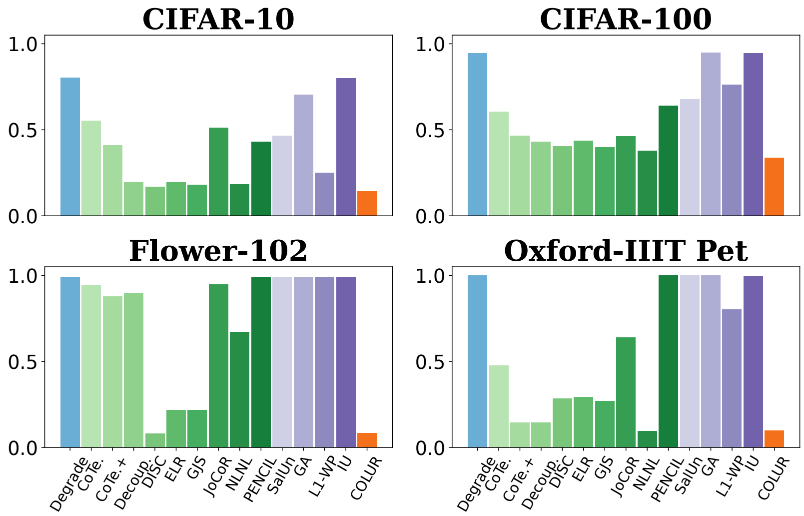


Figure 5: **Error Rates After Label Refinement**. Classification error rates on the noisy label subset $\mathcal{D}_n^u$ of CIFAR-10, CIFAR-100, Flower-102, and Oxford-IIIT Pet datasets with 50% noise ratio. **Degrade** shows degradation on noisy data, while LNL methods, MU methods, and COLUR (highlighted in orange) are all included for comparison. Lower bars indicate better performance.

ods maintain a high true positive rate, as seen by the strong diagonal in their matrices, although some show minor misclassifications. However, COLUR stands out by maintaining a cleaner diagonal with fewer off-diagonal entries, indicating its superiority under minor label noise.

At a high level of noise (90%), most methods experience a significant increase in off-diagonal elements, signaling higher misclassification rates, while COLUR still maintains a relatively strong diagonal with fewer misclassifications than other methods, highlighting its superior ability to correct labels and recover model performance under extreme noise conditions. This performance gap demonstrates the effectiveness of COLUR in mitigating the impact of label noise, surpassing other methods, and achieving a higher true positive rate under extremely high noise levels.

## 4.3 Comparison between LNL and MU for MRR

According to the LUR theory, unlearning can eliminate the impact of false information. Here, we illustrate the efficacy of MU methods, including GA [Graves *et al.*, 2021], IU [Izzo *et al.*, 2021], L1-WP [Liu *et al.*, 2024] and SalUn [Fan *et al.*, 2024], by comparing them with LNL methods. In particular, the ground-truth noisy label datasets $\mathcal{D}_n^u$ are used as the forgetting set for unlearning.

### Quantitative Results

Table 3 shows the performance of all the comparison methods with 50% label noise. By comparing **Original** with **Degrade**, we can find that performance degradation is significant ($\Delta Acc$ is above 20%) after incremental training on $\mathcal{D}^u$. The degradation is more significant on datasets with asymmetric label noise, such as CIFAR-100 and Oxford-IIIT Pet.

LNL methods generally outperform MU methods, although MU methods use the ground-truth noisy label data $\mathcal{D}_n^u$ for unlearning. This is because MU can only mitigate the influence of these noisy label data while no additional corrected knowledge can be incorporated into the degraded models, whereas LNL methods can effectively learn with noisy-label data to restore degraded models with corrected pseudo labels. Especially, **DISC** achieves the overall best performance in all LNL methods thanks to the data augmentation mechanism that improves the robustness to noise.

COLUR takes advantage of both the LNL and the MU methods, so it consistently outperforms the baseline methods across all datasets, achieving results that surpass the **Original** models in most cases. This demonstrates the efficacy of the inspiration mechanism, namely "relearning after unlearning". The unlearning step effectively mitigates the influence of noisy label data recognized by the co-training models. COLUR performs relearning on the refined labels with improved model confidence.

### Error Labeling Rate After MRR

To evaluate the label correction capabilities of each comparison method, Figure 5 demonstrates the error labeling rate on the noisy label subset $\mathcal{D}_n^u$ after performing model restoration on CIFAR-10, CIFAR-100, Flower-102, and Oxford-IIIT Pet. Lower error rates indicate higher capability for label correction and more robustness to noisy labels. From Figure 5, we find that COLUR achieves overall lower error rates compared to the LNL and MU methods in all datasets and noise types. In particular, MU methods underperform other types of methods because MU methods only reverse the influence of noisy label data, but they are incapable of correcting those labels. Therefore, MU cannot be used for MRR independently.

## 4.4 Ablation Study

We conduct ablation studies on the three key modules of the COLUR framework, the unlearning module (UL, cf. Eq (12)), the Mixup module (MP, cf. Eq (17)), and the label smoothing module for relearning (LS, cf. Eqs (20) and (21)) to evaluate their individual contributions and the

| Methods | | CIFAR-10 | CIFAR-100 | Flower-102 | Oxford-IIIT Pet |
|---|---|---|---|---|---|
| **Raw** | Original | 84.85 | 65.16 | 90.08 | 89.26 |
| | Degrade | 65.06 | 38.55 | 66.47 | 53.69 |
| | $\Delta Acc$ (%) | 23.32↓ | 40.84↓ | 26.21↓ | 39.85↓ |
| **LNL** | CoTe. | 45.44 | 39.00 | 64.22 | 58.44 |
| | CoTe.+ | 59.37 | 52.75 | 63.14 | 86.45 |
| | Decoup. | 80.35 | 56.89 | 67.75 | 87.87 |
| | DISC | <u>82.10</u> | 59.34 | <u>88.53</u> | 65.06 |
| | ELR | 80.91 | 57.11 | 75.98 | 63.10 |
| | GJS | 81.65 | 59.38 | 73.73 | 66.58 |
| | JoCoR | 76.32 | 54.97 | 63.43 | 68.19 |
| | NLNL | 82.00 | <u>60.16</u> | 74.71 | <u>89.21</u> |
| | PENCIL | 80.44 | 56.73 | 66.67 | 55.16 |
| **MU** | GA | 69.75 | 38.41 | 66.96 | 67.38 |
| | L1-WP | 76.05 | 35.29 | 67.25 | 36.93 |
| | IU | 65.34 | 38.66 | 66.76 | 63.94 |
| | SalUn | 79.35 | 54.42 | 65.88 | 58.90 |
| **Ours** | **COLUR** | **87.30** | **63.85** | **91.49** | **91.17** |

Table 3: **Comparison between LNL and MU Methods.** In the **Raw** section, **Original** denotes the model originally trained on $\mathcal{D}^0$ and **Degrade** represents the performance after incremental training on $\mathcal{D}^u$. The **LNL** section groups all the LNL methods selected in this study, and the **MU** section groups all the MU methods, where the true noisy label subset $\mathcal{D}_n^u$ with 50% noise ratio is used for unlearning. The last row shows the results of our proposed method, **COLUR**. The best result from each group is highlighted in **bold**, while the second-best one is <u>underlined</u>.

| `UL` | `LS` | `MP` | Flower-102 | Oxford IIIT Pet |
|---|---|---|---|---|
| | | | 66.47 | 53.69 |
| ✓ | | | 67.25 | 69.15 |
| | ✓ | | 83.33 | 86.16 |
| | | ✓ | 84.02 | 89.40 |
| ✓ | | ✓ | 84.31 | 89.67 |
| | ✓ | ✓ | 87.55 | 90.13 |
| ✓ | ✓ | | 85.00 | 89.75 |
| ✓ | ✓ | ✓ | **90.49** | **91.17** |

Table 4: **Ablation Study.** `UL` refers to the unlearning module, `LS` represents the label smoothing module, `MP` refers to the Mixup module. **Degrade** is included for reference in the 1st row. The full model is shown in the last row and highlighted in green.

effectiveness of their combined usage. Experiments are conducted on the Flower-102 dataset with 50% symmetric noise and the Oxford-IIIT Pet dataset with 50% asymmetric noise. The results presented in Table 4 demonstrate the impact of each module and highlight the performance gains achieved through their integration. By comparing the usage of a specific module, we can find that the performance is correspondingly different. The best results are consistently achieved by the full model, which proves that the design of the COLUR scheme is the most effective for MRR.

## 5 Conclusion

This paper addressed the issue of model degradation due to learning over widespread noisy label data. Inspired by the theory of "learning, unlearning and relearning (LUR)", we propose the COLUR framework to address this promising challenge. Following the principle, the proposed COLUR framework iteratively refines the confidences of the model through machine unlearning and relearning, which can effectively restore and refine the degraded task models on noisy label data. Extensive experiments are conducted on four real datasets of various aspects. All results consistently show the superiority of COLUR compared to other SOTA comparison methods in restoring the degraded model and label correction.

## Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Granted No. 62276190).

## Contribution Statement

Zhihao Sui and Liang Hu contributed equally to this work.

## References

[Algan and Ulusoy, 2021] Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021.

[Bao *et al.*, 2025] Guangyin Bao, Qi Zhang, Zixuan Gong, Jialei Zhou, Wei Fan, Kun Yi, Usman Naseem, Liang Hu, and Duoqian Miao. Wills aligner: Multi-subject collaborative brain visual decoding. In *AAAI*, pages 14194–14202. AAAI Press, 2025.

[Becker and Liebig, 2022] Alexander Becker and Thomas Liebig. Evaluating machine unlearning via epistemic uncertainty. *ArXiv preprint arXiv:2208.10836*, 2022.

[Carratino *et al.*, 2022] Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *Journal of Machine Learning Research*, 23(325):1–31, 2022.

[Chen *et al.*, 2024] Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Machine unlearning via null space calibration. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 358–366. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Cheng *et al.*, 2021] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *ICLR*, 2021.

[Di *et al.*, 2024] Zonglin Di, Zhaowei Zhu, Jinghan Jia, Jiancheng Liu, Zafar Takhirov, Bo Jiang, Yuanshun Yao, Sijia Liu, and Yang Liu. Label smoothing improves machine unlearning, 2024.

[Dunlap and Lowenthal, 2011] Joanna C Dunlap and Patrick R Lowenthal. Learning, unlearning, and relearning: Using web 2.0 technologies to support the development of lifelong learning skills. In *E-infrastructures and technologies for lifelong learning: Next generation environments*, pages 292–315. IGI Global, 2011.

[Englesson and Azizpour, 2021] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *NeurIPS*, 34:30284–30297, 2021.

[Fan *et al.*, 2024] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *ICLR*, 2024.

[Ghosh *et al.*, 2017] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017.

[Golatkar *et al.*, 2020] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, pages 9304–9312, 2020.

[Gong *et al.*, 2024] Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, Ke Liu, and Yu Zhang. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. In *NeurIPS*, 2024.

[Graves *et al.*, 2021] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[Izzo *et al.*, 2021] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pages 2008–2016. PMLR, 2021.

[Jiang *et al.*, 2018] 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 *ICML*, volume 80, 2018.

[Karim *et al.*, 2022] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. UNICON: combating label noise through uniform selection and contrastive learning. In *CVPR*, 2022.

[Kim *et al.*, 2019] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *CVPR*, pages 101–110, 2019.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

[Klein, 2008] Emily J Klein. Learning, unlearning, and relearning: Lessons from one school's approach to creating and sustaining learning communities. *Teacher Education Quarterly*, 35(1):79–97, 2008.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, pages 1885–1894. PMLR, 2017.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's Thesis*, 2009.

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

[Li *et al.*, 2021] Junnan Li, Caiming Xiong, and Steven C. H. Hoi. Mopro: Webly supervised learning with momentum prototypes. In *ICLR*, 2021.

[Li *et al.*, 2023] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *CVPR*, pages 24070–24079, 2023.

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

[Liu *et al.*, 2024] Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, PRANAY SHARMA, Sijia Liu, et al. Model sparsity can simplify machine unlearning. *NeurIPS*, 36, 2024.

[Lukasik *et al.*, 2020] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, pages 6448–6458. PMLR, 2020.

[Malach and Shalev-Shwartz, 2017] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *NeurIPS*, 30, 2017.

[Miao *et al.*, 2024] Jiaxing Miao, Liang Hu, Qi Zhang, and Longbing Cao. Graph memory learning: Imitating lifelong remembering and forgetting of brain networks. *CoRR*, abs/2407.19183, 2024.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[Ortego *et al.*, 2021] Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *CVPR*, 2021.

[Parkhi *et al.*, 2012] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[Ramkumar *et al.*, 2023] Vijaya Raghavan T Ramkumar, Elahe Arani, and Bahram Zonooz. Learn, unlearn and relearn: An online learning paradigm for deep neural networks. *Transactions on Machine Learning Research*, 2023.

[Ryan and Frankland, 2022] Tomás J Ryan and Paul W Frankland. Forgetting as a form of adaptive engram cell plasticity. *Nature Reviews Neuroscience*, 23(3):173–186, 2022.

[Sarfraz *et al.*, 2021] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Noisy concurrent training for efficient learning under label noise. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 3159–3168, 2021.

[Sha *et al.*, 2024] Alyssa Shuang Sha, Bernardo Pereira Nunes, and Armin Haller. " forgetting" in machine learning and beyond: A survey. *arXiv preprint arXiv:2405.20620*, 2024.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019.

[Wang *et al.*, 2019] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

[Warnecke *et al.*, 2021] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

[Wei *et al.*, 2020] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13726–13735, 2020.

[Wei *et al.*, 2022] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. In *ICML*, 2022.

[Xu *et al.*, 2019] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, 2019.

[Xu *et al.*, 2023] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), August 2023.

[Yi and Wu, 2019] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, pages 7017–7025, 2019.

[Yu *et al.*, 2019] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, volume 97, 2019.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.

[Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

[Zhang *et al.*, 2021a] Qi Zhang, Longbing Cao, Chongyang Shi, and Liang Hu. Tripartite collaborative filtering with observability and selection for debiasing rating estimation on missing-not-at-random data. In *AAAI*, pages 4671–4678. AAAI Press, 2021.

[Zhang *et al.*, 2021b] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *ICLR*, 2021.

[Zhou *et al.*, 2021] Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *CVPR*, pages 72–81, 2021.