LNL+K: Enhancing Learning with Noisy Labels Through Noise Source Knowledge Integration

Siqi Wang and Bryan A. Plummer

Boston University, Boston MA 02215, USA
{siqiwang,bplum}@bu.edu

Abstract. Learning with noisy labels (LNL) aims to train a high-performing model using a noisy dataset. We observe that noise for a given class often comes from a limited set of categories, yet many LNL methods overlook this. For example, an image mislabeled as a cheetah is more likely a leopard than a hippopotamus due to its visual similarity. Thus, we explore Learning with Noisy Labels with noise source Knowledge integration (LNL+K), which leverages knowledge about likely source(s) of label noise that is often provided in a dataset's meta-data. Integrating noise source knowledge boosts performance even in settings where LNL methods typically fail. For example, LNL+K methods are effective on datasets where noise represents the majority of samples, which breaks a critical premise of most methods developed for LNL. Our LNL+K methods can boost performance even when noise sources are estimated rather than extracted from meta-data. We provide several baseline LNL+K methods that integrate noise source knowledge into state-of-the-art LNL models that are evaluated across six diverse datasets and two types of noise, where we report gains of up to 23% compared to the unadapted methods. Critically, we show that LNL methods fail to generalize on some real-world datasets, even when adapted to integrate noise source knowledge, highlighting the importance of directly exploring LNL+K[1].

Keywords: Learning with Noisy Labels · Dominant Noise · Noise Sources

1 Introduction

High-quality labeled data is valuable for training deep neural networks, but it's costly and often corrupted in real-world datasets [19, 60]. Learning with Noisy Labels (LNL) [36] aims to learn from noisy training data while achieving strong generalization performance [2, 46]. Prior work addresses this task along two main themes: one aligns the noisy data classifier with the clean data classifier through estimated noise transitions [6, 21, 31, 42, 57, 63, 66], while the other discriminates between noisy and clean samples [11, 15, 17, 18, 22, 29, 35, 53]. The core challenge in both types of methods centers on distinguishing potential clean and noisy samples. For example, in Fig. 1-a the input image contains a cat that looks dissimilar to the other cat samples. Thus, prior work would find it challenging to

[1] Code available: https://github.com/SunnySiqi/LNL_K

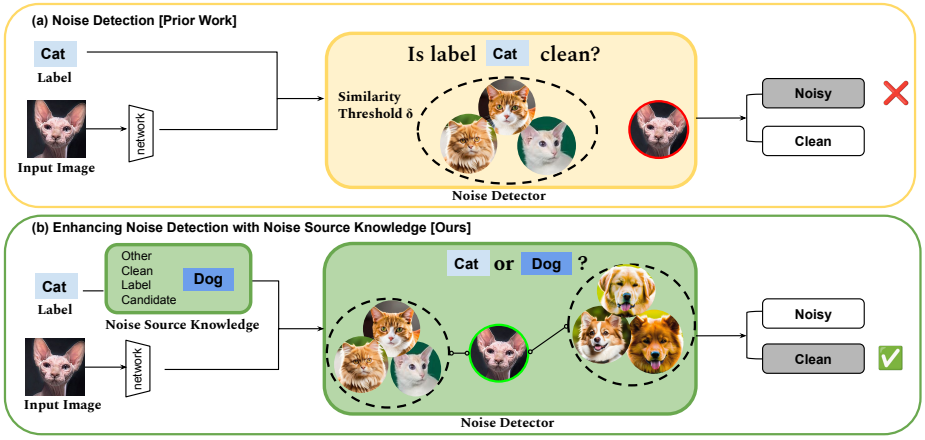

Fig. 1: (Best view in color.) **Comparison of LNL and LNL+K on a hard-negative clean sample**. (a) Traditional LNL methods (*e.g.*, [17, 18, 35, 53]) classify an input image as having a noisy label based on a similarity threshold between the sample and its (majority) class features. (b) In contrast, LNL+K methods identify the sample as a clean label by considering the noise source *dog*. Specifically, since probability of *cat* is higher than that of *dog* and aligns more closely with the *cat* in the feature space, LNL+K judges it as more likely a *cat* image.

identify this sample as clean (*e.g.* [17, 18, 35, 53]). However, as shown in Fig. 1-b, with external knowledge, such as knowing that if the *cat* is mislabeled, it is more likely to be mislabeled as a *dog*, we can identify this sample as a clean label.

We observe that noise source knowledge already exists or can be estimated in real-world datasets. Labels are rarely uniformly corrupted across all classes, and some classes are more easily confused than others [48]. *E.g.*, visually similar objects are often mislabeled: wolf and coyote [45], automobiles and trucks [20]. Furthermore, in scientific settings, certain categories are intentionally designed to establish causality and can be treated as noise sources during training. These are often referred to as a *control i.e.*, *do-nothing* group [54]. For example, cell painting images are labeled with a treatment applied to the cells, but those treatments may have little-to-no effect, meaning that most cells would visually resemble the *control* class (*i.e.*, their true label should be *control*). The noise ratio in this setting can be over 50% [41], which means prior work [17, 18, 35, 53] would be prone to incorrectly consider the more prominent noisy images as the *true* label. Thus, integrating noise source knowledge offers significant potential, particularly in scientific domains with high noise ratios.

To this end, we explore **L**earning with **N**oisy **L**abels through noise source **K**nowledge integration (LNL+K). In contrast to traditional LNL tasks, we assume that we are given some knowledge about noisy label distribution. *i.e.*, noisy labels tend to originate from specific categories (*e.g.*, that the dog is the potential noise source Fig. 1-b). The integration of knowledge about noise sources helps

discriminate clean samples in two ways. First, it aids in the identification of hard negative instances. Second, it enables us to detect noise even at high noise ratios. These two benefits come arise from a shared cause: that our goal is not to identify what labels are clean, but rather what labels are more likely from a noise source. To illustrate, in Fig. 1-b the probability that the input image is a cat is low, but it more like the cat instances than the dogs (the noise source), so it would still be recognized as a clean label. When there are many noisy images, resulting in all cat images producing low probability. LNL+K would separate those that also have high probability of being from a noise source (as noisy) from images with predictions that are higher for non-noise source categories (as clean).

Han *et al*. [10] is the most similar work to ours, which introduces a form of noise supervision by removing invalid noise transitions through human cognition. Thus, while Han *et al*. also aims to leverage some knowledge about the noise source, they focus on estimating noise transitions to avoid overfitting to noisy labels. In our paper, we update and greatly expand on their initial work, including introducing a unified framework with which we can adapt recent methods from LNL to our LNL+K task (*e.g.*, [17, 18, 27, 35, 53]) and investigating new noise settings designed to reflect applications to scientific datasets. Our experiments report up to an 8% gain under asymmetric noise and a remarkable 15% accuracy boost under dominant noise on synthesized noise using CIFAR-10/CIFAR-100 [20]. We also obtain a 1-2% gain on a diverse set of four real-world noisy datasets including two image-based cell profiling datasets [40], CHAMMI-CP [5] and BBBC036 [4], and two natural image datasets, Animal-10N [45] and Clothing1M [58].

In summary, our contributions are:

- We explore an important but overlooked task, termed **LNL+K**: Enhancing **L**earning with **N**oisy **L**abels through noise source **K**nowledge integration. We also design a new noise setting that is widespread in real world: dominant noise, where noisy samples can be the majority of a labeled category distribution.
- We define a unified framework for clean label detection in LNL+K that we use to adapt LNL with noise source knowledge to serve as baselines for LNL+K.
- We analyze the robustness of LNL+K methods on incomplete and noisy knowledge, explore estimating noise source knowledge from noise transition estimation methods, and define *knowledge absorption rate* which can measure how well an LNL method can be transferred to LNL+K tasks, providing an optimizing objective for improving LNL+K methods.

2 Related Work

***LNL with noise supervision* methods** are precursors to the broader concept of LNL+K [10, 12, 28, 52, 65]. *Noise supervision* refers to possessing prior knowledge about the noise present in the dataset. For example, a small clean dataset provides noise source supervision to create more accurate estimates of the noise distribution. However, obtaining human-verified clean datasets is costly and often unavailable. Han *et al*. [10] propose using human cognition of invalid class transitions as *mask* to reduce the burden of transition matrix estimation. However,

this approach is constrained to classifier-consistent methods, and the outcomes become unreliable when the noise structure is misidentified. In contrast to these methods, our LNL+K task does not require complete knowledge. Our goal is to use any existing knowledge about noise sources in a dataset, and we find even partial or noisy noise source knowledge can boost performance.

***Classifier-consistent* methods** align a classifier trained on noisy data with the optimal classifier, which often minimizes errors on clean data. Given input X with true label Y and noisy label $\widetilde{Y}$ the goal is to infer the clean class posterior probability $P(Y|X=x)$ using the noisy class posterior probability $P(\widetilde{Y}|X=x)$ (which can be learned using noisy data) and the transition matrix $T(X=x)$ where $T_{ij}(X=x) = P(\widetilde{Y}=j|Y=i, X=x)$. To estimate the transition matrix, some work uses *anchor points*, samples with very high probability of belonging to a certain class [31,34,38,42]. To avoid using additional clean data, some work focuses on estimating the transition matrix with noisy data [6, 21, 25, 26, 32, 57, 63, 66]. This can be achieved with the density ratio estimation method [51] and matrix decomposition [31, 38, 63], or by training a network to predict the transition matrix [61, 64]. Statistically consistent methods train both noisy and clean data indiscriminately but heavily depend on the accuracy of the noise transition matrix, which becomes particularly challenging with high noise ratios. As we will show, LNL+K can indirectly enhance datasets that require noise source estimation by combining our proposed task with methods for estimating the noise source.

***Classifier-inconsistent* methods** discriminate between clean and noisy labels and handle them differently during training. To discriminate between clean and noisy samples many methods use losses that detect noisy samples with high loss values [1, 16, 22], and some probability-distribution-based approaches select clean samples with high confidence [9, 14, 24, 37, 47, 50]. However, these assumptions may not always hold true, especially with hard negative samples along distribution boundaries. *I.e.*, samples selected by these approaches are more likely *easy* samples instead of *clean* samples. Feature-based approaches utilize the input before the softmax layer – high-dimensional features [18, 35], which are less affected by noisy labels [3, 23, 62]. To differentiate the training of noisy and clean samples, there are methods adjusting the loss function [15, 33, 53, 59, 67], using regularization techniques [13,29,56], multi-round learning only with selected clean samples [8,43,55], and training noisy samples with semi-supervised learning (SSL) techniques [17, 22, 27, 44, 49]. To our knowledge, most statistically inconsistent methods often overlook the valuable resource of noise distribution knowledge in the context of LNL. LNL+K makes a unique contribution by utilizing noise source knowledge to detect clean samples.

3 Learning with Noisy Labels + Knowledge (LNL+K)

Learning with Noisy Label Source Knowledge (LNL+K) aims to find the parameter set θ^* for the classifier f_θ that achieves the highest accuracy on the clean test set when trained on the noisy dataset D **with noise source knowledge** D_{ns}. Suppose we have a dataset $D = \{(x_i, \widetilde{y}_i)_{i=1}^n \in R^d \times K\}$, where $K = \{1, 2, ..., k\}$

is the categorical label for k classes. $(x_i, \widetilde{y_i})$ denotes the $i-th$ example in the dataset, such that x_i is a $d-dimiensional$ input in R^d and $\widetilde{y_i}$ is the label. $\{\widetilde{y_i}\}_{i=1}^n$ might include noisy labels and the true labels $\{y_i\}_{i=1}^n$ are unknown. However, we have some prior knowledge about noisy label sources. This knowledge can take various forms—it can be precise, such as the noise transition matrix obtained through noise modeling methods, or it can be imprecise and incomplete, stemming from human cognition, *e.g.*, classes like *cat* and *lynx* are visually similar and are more likely mislabeled with each other [45]. Additionally, knowledge can be derived from the dataset design, *e.g. control* class serves as the noise source in scientific datasets. Thus, the noise source distribution knowledge D_{ns} can be represented in different ways. One representation is by a probability matrix $P_{k\times k}$, where P_{ij} refers to the probability that a sample in class i is mislabeled as class j. Alternatively, it can also be represented using a set of label pairs $LP = \{(i,j)|i,j \in K\}$, where (i,j) indicates that samples in class i are more likely to be mislabeled as class j. For the convenience of formulating the following equations, noise source knowledge D_{c-ns} represents the set of noise source labels of category c. *I.e.*, $D_{c-ns} = \{i|i \in K \wedge (P_{ic} > 0 \vee (i,c) \in LP)\}$.

3.1 A Unified Framework for Clean Sample Detection with LNL+K

To make our framework general enough to represent different LNL methods, we define a unified logic of clean sample detection. Formally, consider sample x_i with a clean categorical label c, *i.e.*,

$$y_i = c \leftrightarrow \widetilde{y_i} = c \wedge p(c|x_i) > \delta, \tag{1}$$

where $p(c|x_i)$ is the probability of sample x_i with label c and δ is the threshold for the decision. Different methods vary in how they obtain $p(c|x_i)$. For example, loss-based detection uses $Loss(f_\theta(x_i), c)$ to estimate $p(c|x_i)$ [1,16,22], probability-distribution-based methods use the logits or classification probability score $f_\theta(x_i)$ [14,24,37,47,50], and feature-based methods use $p(c|x_i) = M(x_i, \phi_c)$ [18,35], where M is a similarity function, $\phi_c = D(g(X_c))$ is the distribution of features labeled as category c, *i.e.*, $X_c = \{x_i|\widetilde{y_i} = c\}$, $g(X_c) = \{g(x_i, c)|x_i \in X_c\} \sim \phi_c$, and $g(\cdot)$ is a feature mapping function. Feature-based methods often vary in their feature mapping $g(\cdot)$ function and similarity function M.

LNL+K adds knowledge D_{ns} by comparing $p(c|x_i)$ with $p(c_n|x_i)$, where c_n is the noise source label. When category c has multiple noise source labels, $p(c|x_i)$ should be greater than any of them. In other words, the probability that sample x_i has label c (*i.e.*, $p(c|x_i)$), not only depends on its own value, but how it compares to noise sources. *E.g.*, the cat input image in Fig. 1 has low predicted probability of being a cat, *i.e.*, $p(cat|x_i) < \delta$, so LNL methods would identify it as noise (shown in Fig. 1-a). However, for LNL+K in Fig. 1-b, we compare the likelihood of this image being a cat to the probability of belonging to the noise source dog class. Thus, since $p(cat|x_i) > p(dog|x_i)$, it is marked a clean sample.

To summarize, the propositional logic of LNL+K is:

$$y_i = c \leftrightarrow \widetilde{y_i} = c \wedge p(c|x_i) > \max(\{p(c_n|x_i)|c_n \in D_{c-ns}\}). \tag{2}$$

Whereas conventional LNL methods would select examples with the highest probably for a given class, Eq. 2 in LNL+K differs in that:

1. The selected sample's probability may not be its highest predicted category. This is particularly beneficial for identifying hard negative samples, such as images with similar backgrounds. *I.e.*, while the objects themselves may not exhibit similar features and are not considered noise sources, the shared background can cause model confusion for LNL methods.
2. We introduce cross-class comparisons with feature similarity. Most existing LNL methods utilizing feature space similarity do not incorporate cross-class comparisons, focusing solely on the likelihood samples belong to their designated class (Eq. 1). Although AUM [39] introduced cross-class comparisons to the non-assigned label with the highest probability, it was limited to logits and performed poorly at high noise levels like those we study in this paper.

3.2 Incorporating Noise Source Knowledge into LNL methods

We adapt several recent methods from the LNL literature as baselines for our LNL+K task. Our adaptations enhance the detection of clean samples in *inconsistent-classifier methods*. Once the probability of samples being clean is determined, the remainder of the training process follows the original methods. We provide a summary of each adaptation below, but additional details can be found in the supplementary. Algorithm 1 summarizes the steps of our framework, offering a unified approach to integrating noise source knowledge through cross-class comparisons. Note that each base model primarily differs in the function $p(c|x_i)$, which calculates the probability of a sample being clean.

CRUST^{+k} adapts CRUST [35], which uses the pairwise gradient distance within the class for clean sample detection. A clean sample subset is selected with the most similar gradients clustered together. To estimate the likelihood of a sample label being clean in CRUST^{+k}, we apply CRUST on the combined sample set of label class and noise source class. Specifically, for a target label class c, for each noise source c_{ns} in D_{c-ns}, we create the union set of samples $\{x_i|\widetilde{y_i} = c \vee \widetilde{y_i} = c_{ns}\}$. Then CRUST is applied on this union set to identify the clean samples for noise source class c_{ns}. If a sample with label c is selected as part of the clean cluster of c_{ns}, we assume its label is noisy. Let $CRUST(x, c) > 0$ be the indicator that sample x is identified as a clean sample for label c through the computation of the gradient directed towards label c, where x falls within the cluster of similar gradients. Thus, clean c labeled samples identified by CRUST^{+k} are those that satisfy the following: $\{x|CRUST(x, c_{ns}) <= 0, \forall c_{ns} \in D_{c-ns}\}$.

FINE^{+k} is derived from FINE [18], utilizing the alignment between sample and label class features for detection. This alignment is determined by the cosine distance between the sample's features and the eigenvector of the class feature gram matrix, serving as the feature representation for that category. FINE employs a Gaussian Mixture Model (GMM) on the alignment distribution to categorize samples into clean and noisy groups. FINE^{+k} enhances this by incorporating noise source class information into the alignment calculation. In

FINE^{+k}, the clean probability is the difference between the alignment with the label class and the noise-source-class alignment. If $g(x)$ represents the sample feature and $G(c)$ is the feature representation of class c, then FINE fits a GMM directly on $Sim < g(x), G(c) >$ similarity values, while FINE^{+k} fits a GMM on the alignment difference scores between the labeled class and noise source classes. *I.e.*, $Sim < g(x), G(c) > - \max(\{Sim < g(x), G(n_{ns}) > | n_{ns} \in D_{c-ns}\})$.

SFT^{+k} is based on SFT [53], which identifies noisy samples by comparing their predictions over a few recent epochs. A sample is detected as noisy if it used to be classified correctly, but it is misclassified in the latest epoch. SFT^{+k} is adapted by restricting the misclassified labels only to noise source labels.

UNICON^{+k} is adapted from UNICON [17], which estimates the clean probability by using Jensen-Shannon divergence (JSD), a metric for distribution dissimilarity. Disagreement between predicted and one-hot label distributions is utilized, ranging from 0 to 1, with smaller values indicating a higher probability of the label being clean. UNICON^{+k} integrates the noise source knowledge by adding the comparison of JSD with the noise source class. If the sample's predicted distribution aligns more with the noise source, it is considered noisy. For sample x with label $\widetilde{y}$ and noise source knowledge about $\widetilde{y}$ as $D_{\widetilde{y}-ns}$, the clean samples detected by UNICON^{+k} are those belonging to $\{x | JSD(x, \widetilde{y}) < \min(\{JSD(x, c_n) | c_n \in D_{y-ns}\})\}$.

DISC^{+k} adapts DISC [27], which identifies clean and noisy samples based on predictions from two diverse augmentations. Clean samples are selected if the confidences of predictions on weak and strong augmentation images both are over certain thresholds. The clean set is defined as $\{x, \widetilde{y} | p_w(\widetilde{y}, x) > \tau_w(t)\} \cap \{x, \widetilde{y} | p_s(\widetilde{y}, x) > \tau_s(t)\}$, where $p(\widetilde{y}, x)$ represents prediction confidences (w: weak, s: strong), and $\tau(t)$ is the dynamic instance-specific threshold (DIST). The formula for $\tau(t)$ is given by $\tau_x(t) = \lambda\tau(t-1) + (1-\lambda)p_x(t)$, where $p_x(t) = \max(\{p(c; x) | c \in K\})$. In DISC^{+k}, the adaptation involves assigning values to $p_x(t)$ by selecting the largest value among the label class and noise source classes, i.e., $p_x(t) = \max(\{p(c; x) | c \in D_{\widetilde{y}-ns}\} \cup \{p(\widetilde{y}; x)\})$.

DualT+X^{+k} combines noise estimation and noise discrimination methods. DualT [63] is a consistent-classifier method that estimates noise transitions by factorizing the transition matrix into two new matrices that are often easier to estimate compared to the original matrix. Its estimated noise transition matrix can serve as input for any LNL+K method denoted as X^{+k}. We use a straightforward approach to obtain the noise sources from the noise transition matrix: for each class c, we select the class $c_n s$ with the second-highest transition probability (as the highest probability corresponds to the class itself) as the noise source.

4 Experiments

We benchmark two types of noise across six diverse datasets on LNL+K. Each type of noise is evaluated using both synthesized noise from CIFAR-10/CIFAR-100 [20] datasets and two real-world noisy datasets. CIFAR-10/CIFAR-100 [20]

Algorithm 1: Noise Source Integration Algorithm.

Input : Inputs $X = \{x_i\}_{i=1}^n$, noisy labels $Y = \{\widetilde{y}_i\}_{i=1}^n$, probability function p in adaptation-base-method, noise source knowledge D_{ns}

Output : Probabilities of samples being clean $P(X) = \{p(\widetilde{y}_i|x_i)\}_{i=1}^n$

```
for i ← 1 to n do
    p_i ← p(ỹ_i|x_i) ; // Probability of given label ỹ_i being clean.
    for c in D_ns do
        // Loop through noise sources.
        if p(c|x_i) > p_i then
            /* If x_i is more likely to belong to the noise source
               label c, then ỹ_i is considered as the noisy label.   */
            p_i ← 0;
            break;
        end
    end
end
```

dataset contains 10/100 classes with 5,000/500 images per class for training and 1,000/100 images per class for testing, respectively. Synthesized noisy labels are generated based on the noise type for training and validation sets, while ground truth labels are retained for testing and result analysis. For real-world noisy datasets, noise source knowledge is found in the dataset's meta-data (*e.g.* example confusion matrix [58] or experiment design of *controls* as a noise source for cell datasets [4]), *i.e.*, no new annotations are obtained. Additional experimental setup and implementation details can be found in the supplementary.

Baselines. In addition to the baseline methods detailed in Section 3.2, we introduce three additional points of comparison. First, *Baseline* trains on noisy datasets without modifications, *i.e.*, it employs no LNL or LNL+K methods. Second, we include two *noise transition estimation methods*: DualT [63] and GT-T, where GT-T denotes the method trained with a ground truth transition matrix, serving as an upper bound for methods focused on estimating this matrix (*e.g.*, DualT [63]). Note GT-T only applies to synthesized noise experiments due to the absence of ground truth in real-world datasets. The ground truth transition matrix is also used as the noise source knowledge by our $+k$ methods on CIFAR. Third, we compare to SOP [30], a regularization-focused LNL approach.

4.1 Dominant Noise Experiments

Synthesized Dominant Noise and Noisy Cell Image Datasets

– **Dominant Noise** is a novel setting designed to simulate high-noise ratios in real-world settings, particularly in scientific datasets. In this setup, we categorize classes as either *dominant* or *recessive*, where samples mislabeled as *recessive* likely belong to the *dominant* class. *E.g.* in Fig. 2, the noisy samples of *recessive* class A are from *dominant* class B. Our dominant noise simulation can be seen as a special case of unbalanced asymmetric noise with three key

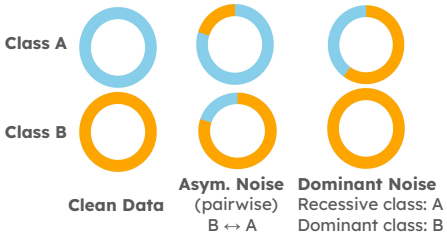

Fig. 2: Comparisons of noise types. Asymmetric noise can occur bidirectionally with limited noise ratios, while Dominant noise can exceed 50% in recessive classes, with clean dominant classes.

features mirroring noise in scientific datasets: 1) The noise ratio can exceed 50%, reflecting scenarios where experiments fail to produce a significant effect, resulting in a high noise ratio [41]. 2) The noise source is always known, as scientists need a *baseline* for their experiments. 3) The noise is one-directional, with the *control* class intentionally devoid of noise sources.
For CIFAR-10/CIFAR-100 [20], we define half the categories as *recessive* and the other half as *dominant*. Noisy labels are generated by labeling images in *dominant* classes as *recessive*, thus *dominant* classes act as noise sources for the *recessive* classes. *Dominant* noise can create a skewed distribution (see example in Fig. 2), which challenges the informative dataset assumption used by prior work [7]. To maintain balance after label corruption, we select different sample numbers from *recessive* and *dominant* classes, see supplementary for detailed noise composition information.

– **Cell Datasets BBBC036 and CHAMMI-CP** contain single U2OS cell (human bone osteosarcoma) images from Cell Painting datasets [4], which represent large treatment screens of chemical and genetic perturbations. Each treatment is tested and then imaged with the Cell Painting protocol [4], which is based on six fluorescent markers captured in five channels. Our goal is to classify the effects of treatments with cell morphology features trained by the model. A significant challenge is that cells react differently to the treatment, *i.e.*, some show minimal differences from controls, creating noisy labels where images look like *controls* (*doing-nothing* group) but are labeled as treatments. Approximately 1,300 of the 1,500 treatments show high similarity with the *controls* [4], suggesting **the majority of the weak-treatment cell images may be noisy**. BBBC036[2] and CHAMMI-CP[3] sampled single-cell images from 1,500 bioactive compounds (treatments), including the *control* group. For BBBC036, we sampled 100 treatments, one of which is labeled as *controls*, resulting in 132,900 training, 14,257 validation, and 22,016 test images. For CHAMMI-CP, we removed three treatments that only appeared in the test set, resulting in four classes, *weak*, *medium*, *strong* treatments, and *control*, with 36,360 training, 3,636 validation, and 13,065 test images. The *control* category is considered the noise sources for all other classes.

[2] Available at `https://bbbc.broadinstitute.org/image_sets`
[3] Available at `https://zenodo.org/record/7988357`

Table 1: Results for dominant noise on CIFAR-10 and CIFAR-100 [20] datasets, along with Cell datasets [4, 5]. The best test accuracy is marked in bold, and the better result between LNL and LNL+K methods are underlined. We find incorporating source knowledge helps in most cases. See Section 4.1 for discussion.

	CIFAR-10 [20]		CIFAR-100 [20]		CHAMMI-CP	BBBC036
Noise ratio	0.5	0.8	0.5	0.8	[5]	[4]
Baseline	85.46±0.25	78.99±0.07	41.41±1.47	27.03±0.12	71.54±0.45	63.49±0.62
DualT [63]	83.70±0.04	46.96±0.07	27.04±0.07	19.94±0.04	70.73±0.17	61.54±0.61
GT-T	85.24±0.06	76.03±0.04	48.39±0.21	35.96±0.04	-	-
SOP [30]	86.94±0.37	80.65±0.71	55.78±0.68	45.94±0.62	77.55±0.23	60.94±0.38
CRUST [35]	80.46±0.17	65.79±0.62	48.87±0.31	35.56±1.38	78.02±0.31	63.06±0.65
CRUST^{+k}	87.19±0.08	80.54±0.30	51.56±0.31	38.07±2.05	**79.81**±0.56	**65.07**±0.71
FINE [18]	84.43±0.38	75.45±0.74	52.87±0.98	39.45±0.25	67.27±0.82	56.80±0.87
FINE^{+k}	88.00±0.11	80.52±0.28	54.77±1.68	42.25±0.27	67.02±0.73	57.01±0.40
SFT [53]	85.43±0.13	75.43±0.12	48.21±1.21	41.76±1.34	76.08±0.25	51.71±0.82
SFT^{+k}	87.31±0.15	76.78±0.38	51.21±1.14	37.96±0.05	77.75±0.42	59.18±1.33
UNICON [17]	88.43±0.14	81.37±0.43	57.92±0.43	42.70±0.50	71.45±0.03	33.98±1.03
UNICON^{+k}	89.21±0.42	82.27±0.29	61.55±0.13	48.47±0.40	71.04±0.14	42.17±0.31
DISC [27]	91.58±0.21	85.89±0.16	64.97±0.17	49.79±0.20	74.04±0.11	40.55±0.18
DISC^{+k}	**91.88**±0.15	**86.70**±0.03	**65.96**±0.15	**50.74**±0.11	75.38±0.30	63.32±0.49

Results. Table 1 reports classification accuracy in dominant noise setting, where LNL+K methods report the best performance in all cases. We see that as the noise ratio is increased, the gains reported by LNL+K methods also increase. In the synthetic noise scenarios, there is an average improvement of 3% in the 80% noise ratio, with CRUST^{+k} achieving an impressive 15% improvement on CIFAR-10. For the cell datasets, high feature similarity between certain treatments and the *control* group can lead to significantly high noise ratios, strongly influencing the class distribution. On BBBC036, DISC^{+k} achieves a significant 23% boost, but this only brings the method in line with standard training. Notably, only CRUST^{+k} outperforms standard training, boosting top-1 accuracy by 1.5%, whereas all LNL methods fail to improve performance. This helps illustrate the importance of exploring LNL+K. We also note that in both cell datasets there are additional sources of noise that are not addressed by LNL+K. Specifically, there are technical variations that arise due to differences in the experimental protocol used to collect the images [40]. Thus, our noise source knowledge should be considered incomplete, yet still provide enough signal to boost performance.

To understand the source of our performance gains, in Table 2 we report the effect integrating knowledge in LNL+K has on identifying clean samples. We find that our LNL+K boosts sample selection accuracy across all methods, with the largest gain being an improvement in recall by over 60% for CRUST while also improving precision by 15%. Due to this remarkable gain in sample selection

Table 2: Results for 0.8 dominant noise on CIFAR-10 [20]. We report precision and recall for the selected *clean* subset and the model's classification accuracy. Integrating knowledge (+K) significantly enhances sample selection performance. See Section 4.1 for further discussion.

	Precision	Recall	Cls. Acc.
CRUST [35]	72.29	36.15	65.79
CRUST^{+k}	**87.67**	**99.04**	**80.54**
FINE [18]	88.53	61.89	75.45
FINE^{+k}	**89.64**	**99.61**	**80.52**
SFT [53]	97.27	94.37	75.43
SFT^{+k}	**98.99**	**94.95**	**76.78**

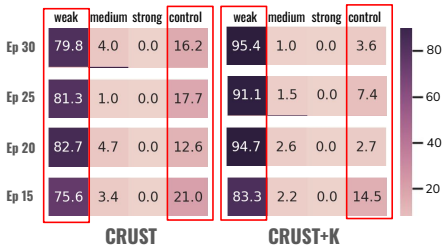

Fig. 3: Class prediction confusion matrix for **weak treatment** cell images in the CHAMMI-CP [5] dataset, normalized to sum to 100%. The integration of knowledge (+K) enhances the method's capability to distinguish weak treatment from a high-ratio *control* class. See Section 4.1 for more details.

accuracy, CRUST^{+k} performs on par or better than FINE^{+k} and SFT^{+k} on cell datasets, eliminating the apparent advantage these methods had over CRUST in the LNL setting. To show the effect that knowledge has on noisy class predictions during training in cell datasets, Fig. 3 reports the predictions made on weak treatment cells (*i.e.*, the class with the highest noise) on CHAMMI-CP's test set during training with CRUST. We find that including knowledge enables our approach to more effectively identify these treatments.

4.2 Asymmetric Noise Experiments

Synthesized Asymmetric Noise and Noisy Online Image Datasets

- **Asymmetric Noise** simulates real-world scenarios where visually similar objects are mislabeled as each other. *I.e.*, labels are corrupted for visually similar classes, such as trucks ↔ automobiles in CIFAR-10 [20]. Asymmetric noise is synthesized in a **pair-wise** manner, *i.e.*, $P(A \rightarrow B) = P(B \rightarrow A)$. As this noise is bidirectional, when the noise ratio surpasses 50%, the model lacks cues to discern the true label from the minority class, therefore, our experiments focus on two noise ratios: 20% and 40%. The supplementary contains a list of confusing pairs, which serve as noise sources in our experiments.
- **Online Image Datasets Animal-10N and Clothing1M** contain images that were obtained by crawling several online search engines using predefined labels as search keywords or extracting noisy labels from surrounding text. The noisy labeled data is utilized for training and validation, while a smaller test set has manually annotated clean labels. Animall-10N [45] has 10 confusing animal classes with a total of 50,000 training images and 5000 testing images. The dataset author noted 5 pairs of classes that can be easily confused[4],

[4] Available at `https://dm.kaist.ac.kr/datasets/animal-10n/`

Table 3: Results for asymmetric noise on CIFAR-10 and CIFAR-100 [20] datasets, along with Animal-10N [45]. The best test accuracy is marked in bold, and the better result between LNL and LNL+K methods are underlined. We find knowledge-adapted methods can alter the rankings of the base methods. (*e.g.* SFT and FINE at a noise ratio of 0.4 on CIFAR-100.) See Section 4.2 for discussion.

	CIFAR-10 [20]		CIFAR-100 [20]		Animal-10N
Noise ratio	0.2	0.4	0.2	0.4	[45]
Baseline	86.12±0.42	77.18±0.30	62.96±0.12	59.07±0.08	80.32±0.20
DualT [63]	92.24±0.10	66.23±0.03	53.61±1.49	52.03±1.92	81.14±0.28
GT-T	92.51±0.03	89.68±0.13	73.88±0.04	66.61±0.03	-
SOP [30]	92.85±0.49	89.93±0.25	72.60±0.70	70.58±0.30	83.93±0.35
CRUST [35]	91.94±0.05	89.40±0.03	60.75±1.87	59.79±0.89	81.88±0.13
CRUST^{+k}	89.47±0.17	84.96±0.91	62.44±0.84	61.07±0.16	81.74±0.08
FINE [18]	89.07±0.03	85.51±0.18	65.42±0.11	65.11±0.11	81.15±0.11
FINE^{+k}	90.87±0.04	89.15±0.26	73.59±0.12	72.87±0.11	82.27±0.10
SFT [53]	92.67±0.04	89.77±0.14	74.41±0.05	69.51±0.06	82.24±0.10
SFT^{+k}	93.19±0.08	90.55±0.06	74.29±0.14	70.94±0.13	82.88±0.18
UNICON [17]	92.42±0.04	91.51±0.12	75.95±0.04	73.08±0.07	87.76±0.06
UNICON^{+k}	92.60±0.07	91.35±0.24	76.87±0.24	73.97±0.11	**88.28**±0.29
DISC [27]	94.82±0.04	93.24±0.04	76.02±0.15	74.36±0.16	86.44±0.14
DISC^{+k}	**95.40**±0.08	**94.05**±0.07	**77.13**±0.05	**75.50**±0.08	86.90±0.10

which serve as noise sources in our experiments. Clothing1M [58] contains approximately one million clothing images. The dataset encompasses 14 classes, with an estimated overall label accuracy of around 60%. However, the label noise is imbalanced, with certain classes experiencing noise levels as high as 80% (*e.g.*, mislabeling *sweater* as *knitwear*). We provide a summary of the noise sources for each class in the supplementary, extracted from the confusion matrix presented in the original paper [58]. Additional details on the integrated noise source knowledge in our experiments is also in the supplementary.

Results. Table 3 summarizes classification accuracy in asymmetric noise settings, highlighting the advantage of LNL+K in visually similar noise cases. Our adaptation methods consistently outperform the original methods across most noise settings. Notably, FINE^{+k} exhibits significant performance improvement, achieving up to an 8% increase in accuracy compared to the base FINE method on CIFAR-100. For CIFAR-100 with a 0.4 noise ratio, the base model SFT achieves 4% higher accuracy than FINE. However, with the integration of knowledge, FINE^{+k} surpasses the performance of SFT^{+k} by 2%. These results underscore the importance of investigating LNL+K tasks. The reported gains on Animal-10N [45] in Table 3 and Clothing1M [58] in Table 4 further illustrate the advantages of integrating knowledge with LNL+K on general, real-world LNL benchmarks.

Table 4: Results on Clothing1M [58] dataset. Results with * are from the referenced paper, others are our implementation. See Section 4.2 for discussion.

Baseline	DivideMix* [22]	ELR* [29]	CORES^{2}* [7]	SOP* [30]	UNICON [17]	UNICON^{+k} (ours)	DISC [27]	DISC^{+k} (ours)
69.45	74.76	72.87	73.24	73.50	74.56	**75.13**	73.30	73.87

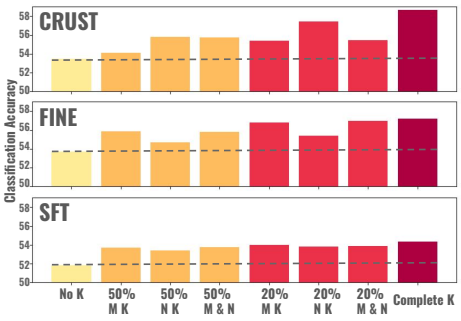

Fig. 4: Comparisions of knowledge-adaptive methods with different degrees of noisy noise sources. (MK: missing knowledge, NK: noisy knowledge and M&N: the combination of these two.) Note: complete knowledge has 50 noise sources. See Section 4.3 for discussion.

4.3 Discussion

Incomplete or noisy knowledge. Noise sources need not be strictly complete or clean to provide benefits. *E.g.*, Fig. 4 shows results of missing and incorrect noise sources on CIFAR-100 with a 20% dominant noise setting. Each recessive class has noise from 50 dominant classes. Missing knowledge (MK): 50% MK means 25 noise sources are missing. Noisy knowledge (NK): 50% NK means 25 noise sources are incorrect. M&N: 50% M&N means 25 correct noise sources. We find partial or incomplete knowledge (columns 2-7) is better than no knowledge (column 1) and sometimes only minorly impacts full knowledge (last column).

Learning noise source knowledge from noise transition estimation methods. We also explored estimating noise source knowledge using DualT [63] with knowledge-adapted methods. Table 5 reports performance, where we find combining DualT with our LNL+K methods boosts performance. Notably, when compared to the original LNL variants from Table 3, our LNL+K models obtain similar or better performance even with estimated noise source knowledge, further validating the importance of our work. Noise source knowledge acts as a bridge between noise estimation and detection methods, enabling knowledge-integrated approaches to function even in the absence of prior information.

Knowledge absorption rate varies for different methods at the same noise settings. From the results in Section 4.1 and Section 4.2, we notice that the accuracy improvements of the adaptation methods vary in different noise settings and methods. We define this different degree of improvement as *knowledge absorption rate*. Considering the unified framework of detecting clean labels in Section 3, $p(c|x_i)$ and $p(c_n|x_i)$ are important factors to *Knowledge absorption rate*. Our baseline methods represent five different methods of estimating $p(c|x_i)$. The

Table 5: Results of combining noise estimation and noise detection algorithms with knowledge on multiple datasets. We find that using noise estimation with LNL+K can still boost performance. See Section 4.3 for discussion.

	CIFAR-10 [20]		CIFAR-100 [20]		CHAMMI-CP	Animal-10N
Asym Noise Ratio	0.2	0.4	0.2	0.4	[5]	[45]
FINE [18]	89.07	85.51	65.42	65.11	67.27	81.15
DualT [63]+FINE^{+k}	89.89	88.87	66.36	62.80	**70.70**	81.84
FINE^{+k}	**90.87**	**89.15**	**73.59**	**72.87**	67.02	**82.27**

results conclude that noise source knowledge might be more helpful to the feature-based clean sample detection methods in high noise ratios. CRUST^{+k} has better performance than FINE^{+k} on high noise ratios in the cell datasets in Table 1. A similar observation is made in dominant noise settings, such as the those explored in Table 2, where CRUST^{+k} received a larger boost to performance than other methods, such as FINE^{+k}. One possible explanation for this is that $p(c|x_i)$ for FINE^{+k} depends on the category feature distribution while CRUST^{+k} focuses on the feature of a single sample and aims to find the subset with minimum gradient distance sum. In other words, when the noise ratio is high, category feature distribution in FINE^{+k} might be skewed while CRUST^{+k} is less affected by finding the optimal cluster. *Knowledge absorption rate* indicates how well an LNL method can transfer to the LNL+K task with noise distribution knowledge, exploring ways to enhance the transferability of LNL methods and optimizing this value are important areas for further investigation.

5 Conclusion

This paper introduces a new task, LNL+K, which leverages noise source distribution knowledge when learning with noisy labels. This knowledge is not only beneficial to distinguish clean samples that are ambiguous or out-of-distribution but also necessary when the noise ratio is so high that the noisy samples dominate the class distribution. Instead of comparing the *similarity* of the samples within the same class to detect the clean ones, LNL+K utilizes the *dissimilarity* between the sample and the noise source for detection. We provide a unified framework of clean sample detection for LNL+K which we use to adapt state-of-the-art LNL methods, CRUST^{+k}, FINE^{+k}, SFT^{+k}, UNICON^{+k} and DISC^{+k} to our task. To create a more realistic simulation of high-noise-ratio settings, we introduce a novel noise setting called *dominant noise*. Results show LNL+K methods have up to 8% accuracy gains over asymmetric noise and up to 15% accuracy gains in the dominant noise setting. Finally, we discuss the robustness of LNL+K towards incomplete and noisy source knowledge and learning source knowledge from noise estimation methods. We also define *knowledge absorption*, which notes the ranking of LNL methods to our task varies from their LNL performance, indicating that direct investigation of LNL+K is necessary.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. DBI-2134696. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: International conference on machine learning. pp. 312–321. PMLR (2019)
2. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., Lacoste-Julien, S.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242 (2017)
3. Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., Liu, T.: Understanding and improving early stopping for learning with noisy labels. Advances in Neural Information Processing Systems **34**, 24392–24403 (2021)
4. Bray, M.A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C., Carpenter, A.E.: Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature protocols **11**(9), 1757–1774 (2016)
5. Chen, Z., Pham, C., Wang, S., Doron, M., Moshkov, N., Plummer, B.A., Caicedo, J.C.: Chammi: A benchmark for channel-adaptive models in microscopy imaging. arXiv preprint arXiv:2310.19224 (2023)
6. Cheng, D., Liu, T., Ning, Y., Wang, N., Han, B., Niu, G., Gao, X., Sugiyama, M.: Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16630–16639 (2022)
7. Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., Liu, Y.: Learning with instance-dependent label noise: A sample sieve approach. arXiv preprint arXiv:2010.02347 (2020)
8. Cordeiro, F.R., Sachdeva, R., Belagiannis, V., Reid, I., Carneiro, G.: Longremix: Robust learning with high confidence samples in a noisy label environment. Pattern Recognition **133**, 109013 (2023)
9. Feng, C., Tzimiropoulos, G., Patras, I.: Ssr: An efficient and robust framework for learning with unknown label noise. arXiv preprint arXiv:2111.11288 (2021)
10. Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., Sugiyama, M.: Masking: A new perspective of noisy supervision. Advances in neural information processing systems **31** (2018)
11. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)
12. Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K.: Using trusted data to train deep networks on labels corrupted by severe noise. In: Advances in neural information processing systems (2018)
13. Hu, W., Li, Z., Yu, D.: Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. arXiv preprint arXiv:1905.11368 (2019)

14. Hu, W., Zhao, Q., Huang, Y., Zhang, F.: P-diff: Learning classifier with noisy labels based on probability difference distributions. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1882–1889. IEEE (2021)
15. Iscen, A., Valmadre, J., Arnab, A., Schmid, C.: Learning with neighbor consistency for noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4672–4681 (2022)
16. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International conference on machine learning. pp. 2304–2313. PMLR (2018)
17. Karim, N., Rizve, M.N., Rahnavard, N., Mian, A., Shah, M.: 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)
18. Kim, T., Ko, J., Choi, J., Yun, S.Y., et al.: Fine samples for learning with noisy labels. Advances in Neural Information Processing Systems **34**, 24137–24149 (2021)
19. Krishna, R.A., Hata, K., Chen, S., Kravitz, J., Shamma, D.A., Fei-Fei, L., Bernstein, M.S.: Embracing error to enable rapid crowdsourcing. In: Proceedings of the 2016 CHI conference on human factors in computing systems. pp. 3167–3179 (2016)
20. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
21. Kye, S.M., Choi, K., Yi, J., Chang, B.: Learning with noisy labels by efficient transition matrix estimation to combat label miscorrection. In: European Conference on Computer Vision. pp. 717–738. Springer (2022)
22. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
23. Li, M., Soltanolkotabi, M., Oymak, S.: Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In: International conference on artificial intelligence and statistics. pp. 4313–4324. PMLR (2020)
24. Li, S., Xia, X., Ge, S., Liu, T.: Selective-supervised contrastive learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 316–325 (2022)
25. Li, S., Xia, X., Zhang, H., Zhan, Y., Ge, S., Liu, T.: Estimating noise transition matrix with label correlations for noisy multi-label learning. Advances in Neural Information Processing Systems **35**, 24184–24198 (2022)
26. Li, X., Liu, T., Han, B., Niu, G., Sugiyama, M.: Provably end-to-end label-noise learning without anchor points. In: International conference on machine learning. pp. 6403–6413. PMLR (2021)
27. Li, Y., Han, H., Shan, S., Chen, X.: 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)
28. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: Proceedings of the IEEE international conference on computer vision. pp. 1910–1918 (2017)
29. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems **33**, 20331–20342 (2020)
30. Liu, S., Zhu, Z., Qu, Q., You, C.: Robust training under label noise by over-parameterization. In: Proceedings of the 39th International Conference on Machine Learning (2022)
31. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence **38**(3), 447–461 (2015)

32. Liu, Y., Cheng, H., Zhang, K.: Identifiability of label noise transition matrix. In: International Conference on Machine Learning. pp. 21475–21496. PMLR (2023)
33. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., Bailey, J.: Normalized loss functions for deep learning with noisy labels. In: International conference on machine learning. pp. 6543–6553. PMLR (2020)
34. Menon, A., Van Rooyen, B., Ong, C.S., Williamson, B.: Learning from corrupted binary labels via class-probability estimation. In: International conference on machine learning. pp. 125–134. PMLR (2015)
35. Mirzasoleiman, B., Cao, K., Leskovec, J.: Coresets for robust training of deep neural networks against noisy labels. Advances in Neural Information Processing Systems **33**, 11465–11477 (2020)
36. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. Advances in neural information processing systems **26** (2013)
37. Nguyen, D.T., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Beggel, L., Brox, T.: Self: Learning to filter noisy labels with self-ensembling. arXiv preprint arXiv:1910.01842 (2019)
38. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1944–1952 (2017)
39. Pleiss, G., Zhang, T., Elenberg, E., Weinberger, K.Q.: Identifying mislabeled data using the area under the margin ranking. Advances in Neural Information Processing Systems **33**, 17044–17056 (2020)
40. Pratapa, A., Doron, M., Caicedo, J.C.: Image-based cell phenotyping with deep learning. Current Opinion in Chemical Biology **65**, 9–17 (2021)
41. Rohban, M.H., Singh, S., Wu, X., Berthet, J.B., Bray, M.A., Shrestha, Y., Varelas, X., Boehm, J.S., Carpenter, A.E.: Systematic morphological profiling of human gene and allele function via cell painting. Elife **6**, e24060 (2017)
42. Scott, C.: A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In: Artificial Intelligence and Statistics. pp. 838–846. PMLR (2015)
43. Shen, Y., Sanghavi, S.: Learning with bad training data via iterative trimmed loss minimization. In: International Conference on Machine Learning. pp. 5739–5748. PMLR (2019)
44. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
45. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International Conference on Machine Learning. pp. 5907–5915. PMLR (2019)
46. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)
47. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5552–5560 (2018)
48. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11244–11253 (2019)

49. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
50. Torkzadehmahani, R., Nasirigerdeh, R., Rueckert, D., Kaissis, G.: Label noise-robust learning using a confidence-based sieving strategy. arXiv preprint arXiv:2210.05330 (2022)
51. Vapnik, V., Braga, I., Izmailov, R.: Constructive setting of the density ratio estimation problem and its rigorous solution. arXiv preprint arXiv:1306.0407 (2013)
52. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 839–847 (2017)
53. Wei, Q., Sun, H., Lu, X., Yin, Y.: Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX. pp. 516–532. Springer (2022)
54. Wikipedia contributors: Treatment and control groups — Wikipedia, the free encyclopedia (2022), `https://en.wikipedia.org/w/index.php?title=Treatment_and_control_groups&oldid=1110767032`, [Online; accessed 24-April-2023]
55. Wu, P., Zheng, S., Goswami, M., Metaxas, D., Chen, C.: A topological filter for learning with label noise. Advances in neural information processing systems **33**, 21382–21393 (2020)
56. Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., Chang, Y.: Robust early-learning: Hindering the memorization of noisy labels. In: International conference on learning representations (2021)
57. Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., Sugiyama, M.: Are anchor points really indispensable in label-noise learning? Advances in neural information processing systems **32** (2019)
58. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2691–2699 (2015)
59. Xu, Y., Cao, P., Kong, Y., Wang, Y.: L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. Advances in neural information processing systems **32** (2019)
60. Yan, Y., Rosales, R., Fung, G., Subramanian, R., Dy, J.: Learning from multiple annotators with varying expertise. Machine learning **95**, 291–327 (2014)
61. Yang, S., Yang, E., Han, B., Liu, Y., Xu, M., Niu, G., Liu, T.: Estimating instance-dependent bayes-label transition matrix using a deep neural network. In: International Conference on Machine Learning. pp. 25302–25312. PMLR (2022)
62. Yao, Q., Yang, H., Han, B., Niu, G., Kwok, J.T.Y.: Searching to exploit memorization effect in learning with noisy labels. In: International Conference on Machine Learning. pp. 10789–10798. PMLR (2020)
63. Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., Sugiyama, M.: Dual t: Reducing estimation error for transition matrix in label-noise learning. Advances in neural information processing systems **33**, 7260–7271 (2020)
64. Yong, L., Pi, R., Zhang, W., Xia, X., Gao, J., Zhou, X., Liu, T., Han, B.: A holistic view of label noise transition matrix in deep learning and beyond. In: The Eleventh International Conference on Learning Representations (2022)
65. Yu, C., Ma, X., Liu, W.: Delving into noisy label detection with clean data. In: Proceedings of the 40th International Conference on Machine Learning (23–29 Jul 2023)

66. Zhang, Y., Niu, G., Sugiyama, M.: Learning noise transition matrix from only noisy labels via total variation regularization. In: International Conference on Machine Learning. pp. 12501–12512. PMLR (2021)
67. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)