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

Siqi Wang and Bryan A. Plummer

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

1 LNL+K baseline methods

For the convenience of formulating the following equations, recall the notations we defined in Section 3 in the main paper. 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. $\{\widetilde{y_i}\}_{i=1}^n$ might include noisy labels and the true labels $\{y_i\}_{i=1}^n$ are unknown. Noise transition probability matrix $P_{k\times k}$, where P_{ij} refers to the probability that a sample in class i is mislabeled as class j. 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. noise source knowledge D_{c-ns} represents the set of noise source labels of category c. *I.e.*, $D_{c-ns} = \{i|i \in K \wedge (P_{ic} > 0 \vee (i,c) \in LP)\}$.

1.1 CRUST^{+k}

The key idea of CRUST [11] is from the neural network Jacobian matrix containing all its first-order partial derivatives. It is proved in their work that the neural network has a low-rank Jacobian matrix for clean samples. In other words, data points with clean labels in the same class often have similar gradients clustered closely together. CRUST [11] is a feature-based method and this approach can be summarized with settings in Section 3.1. The feature used for selection is the pairwise gradient distance within the class: $g(X_c) = \{d_{x_i x_j}(\mathcal{W})|x_i, x_j \in X_c\}$, where $d_{x_i x_j}(\mathcal{W}) = \|\nabla L(\mathcal{W}, x_i) - \nabla L(\mathcal{W}, x_j)\|_2$, $\mathcal{W}$ is the network parameters and $L(\mathcal{W}, x_i) = \frac{1}{2}\sum_{x_i \in D}(y_i - f_\theta(\mathcal{W}, x_i))^2$. CRUST [11] needs an additional parameter β to control the size of the clean selection set X_c'. Given β, the sample x_i is selected as clean if $\|X_c'\| = \beta$ ($\|X_c'\|$ is the size of set X_c') and $x_i \in X_c'$, where $\sum g(X_c')$ has the minimum value. *i.e.*, the selected clean subset X_c' has the most similar gradients clustered together. Thus, we can summarize the similarity metric M for $p(c|x_i)$ as:

$$\begin{aligned} M(x_i, \phi_c, \beta) = 1 \leftrightarrow \exists X_c' \subset X_c \wedge \|X_c'\| = \beta, \\ s.t.\ x_i \in X_c' \wedge (\forall \|X_c''\| = \beta \wedge X_c'' \subset X_c, \\ \sum g(X_c') \leq \sum g(X_c'')), \end{aligned} \tag{1}$$

otherwise $M(x_i, \phi_c, \beta) = 0$. Thus, we can get the propositional logic of CRUST:

$$y_i = c \leftrightarrow \widetilde{y_i} = c \wedge p(c|x_i) = 1 \leftrightarrow M(x_i, \phi_{\widetilde{y_i}}, \beta) = 1. \tag{2}$$

To adapt CRUST to CRUST^{+k} with noise source distribution knowledge. from Eq.2 in the main paper, we have

$$\begin{aligned} \widetilde{y_i} = c \wedge y_i \neq c &\leftrightarrow p(c|x_i) \leq \max(\{p(c_n|x_i)|c_n \in D_{c-ns}\}) \\ &\leftrightarrow \exists c_n \in D_{c-ns}\ s.t.\ p(c_n|x_i) \geq p(c|x_i) \\ &\leftrightarrow \exists c_n \in D_{c-ns}\ s.t.\ p(c_n|x_i) = 1. \end{aligned} \tag{3}$$

To get $p(c_n|x_i)$, we first mix x_i with all the samples in X_{c_n}, *i.e.*, $X_{c_n+} = \{x_i\} \cup X_{c_n}$. Then apply CRUST on this mix set, *i.e.*, calculate the loss towards label c_n and select the clean subset X'_{c_n+}. if $x_i \in X'_{c_n+}$, then $p(c_n|x_i) = 1$. Here is the formulation of CRUST^{+k}, we modify $L(\mathcal{W}, x_i)$ to $L(\mathcal{W}, x_i, c) = \frac{1}{2}\sum_{x_i \in D}(c - f_\theta(\mathcal{W}, x_i))^2$, where we calculate the loss to any certain categories, not limited to the loss towards the label. Similarly, we have $d_{x_i x_j}(\mathbf{W}, c) = \|\nabla L(\mathcal{W}, x_i, c) - \nabla L(\mathcal{W}, x_j, c)\|_2$, $g(X_{c_n+}, c_n) = \{d_{x_i x_j}(\mathbf{W}, c_n)|x_i, x_j \in X_{c_n+}\}$. We use γ to represent the subset size of X_{c+c_n}, which is decided by β and noise source distribution. Finally, we get the similarity metric $M(x_i, \phi_{c_n+}, \gamma)$ as:

$$\begin{aligned} M(x_i, \phi_{c_n+}, \gamma) = 1 \leftrightarrow \exists X'_{c_n+} \subset X_{c_n+} \wedge \|X'_{c_n+}\| = \gamma, \\ s.t.\ x_i \in X'_{c_n+} \wedge (\forall \|X''_{c_n+}\| = \gamma \wedge X''_{c_n+} \subset X_{c_n+}, \\ \sum g(X'_{c_n+}, c_n) \leq \sum g(X''_{c_n+}, c_n)), \end{aligned} \tag{4}$$

otherwise $M(x_i, \phi_{c_n+}, \gamma) = 0$. Combining Eq.2 in the main paper, Eq.2, and Eq.4, $p(c|x_i)$ of CRUST^{+k} method is:

$$\begin{aligned} y_i = c \leftrightarrow \widetilde{y_i} = c \wedge (\forall c_n \in D_{c-ns}, p(c_n|x_i) < p(c|x_i)) \\ \leftrightarrow \widetilde{y_i} = c \wedge (\forall c_n \in D_{c-ns}, p(c_n|x_i) = 0) \\ \leftrightarrow \widetilde{y_i} = c \wedge (\forall c_n \in D_{c-ns}, M(x_i, \phi_{c_n+}, \gamma) = 0). \end{aligned} \tag{5}$$

1.2 FINE^{+k}

Filtering Noisy instances via their Eigenvectors(*FINE*) [6] selects clean samples with the feature-based method. Let $f_{\theta^*}(x_i)$ be the feature extractor output and Σ_c be the gram matrix of all features labeled as category c. The alignment is defined as the cosine distance between feature $\overrightarrow{f_{\theta^*}(x_i)}$ and $\overrightarrow{c}$, which is the eigenvector of the Σ_c and can be treated as the feature representation of category c. FINE fits a Gaussian Mixture Model (GMM) on the alignment distribution to divide samples to clean and noisy groups - the clean group has a larger mean value, which refers to a better alignment with the category feature representation. In summary, feature mapping function $g(x_i, c) = < \overrightarrow{f_{\theta^*}(x_i)}, \overrightarrow{c} >$, and mixture of Gaussian distributions $\phi_c = \mathcal{N}_{clean} + \mathcal{N}_{noisy} = \mathcal{N}(\mu_{g(X_{c-clean})}, \sigma_{g(X_{c-clean})}) +$

$\mathcal{N}(\mu_{g(X_{c-noisy})}, \sigma_{g(X_{c-noisy})})$, where $\mu_{g(X_{c-clean})} > \mu_{g(X_{c-noisy})}$. The similarity metric

$$M(x_i, \phi_c) = \begin{cases} 1 & : \quad \mathcal{N}_{clean}(g(x_i, c)) > \mathcal{N}_{noisy}(g(x_i, c)) \\ 0 & : \quad \mathcal{N}_{clean}(g(x_i, c)) \leq \mathcal{N}_{noisy}(g(x_i, c)). \end{cases}$$

Thus, we have

$$y_i = c \leftrightarrow \widetilde{y}_i = c \wedge p(c|x_i) = 1 \leftrightarrow M(x_i, \phi_{\widetilde{y}_i}) = 1. \tag{6}$$

Next, we show our design of FINE^{+k} with noise source distribution knowledge. The key difference between FINE and FINE^{+k} is that we use the alignment score of the noise source class. For a formal description of FINE^{+k}, We define $g_k(x_i, c, c_n) = g(x_i, c) - g(x_i, c_n)$. Similar to FINE, FINE^{+k} fits a GMM on $g_k(X_c, c, c_n)$, so we have $g_k(X_c, c, c_n) \sim \phi_{k-\{c+c_n\}} = \mathcal{N}_{close-c} + \mathcal{N}_{close-c_n}$, where $\mu_{close-c} > \mu_{close-c_n}$. This can be interpreted in the following way: Samples aligning better with category c should have larger $g(x_i, c)$ values and smaller $g(x_i, c_n)$ values according to the assumption, thus the greater the $g_k(x_i, c, c_n)$, the closer to category c, vice versa, the smaller the $g_k(x_i, c, c_n)$, the closer to category c_n. Then we have

$$\begin{aligned} & M(x_i, \phi_{k-\{c+c_n\}}) = 1 \\ \leftrightarrow\ & \mathcal{N}_{close-c}(g_k(x_i, c, c_n)) > \mathcal{N}_{close-c_n}(g_k(x_i, c, c_n)) \end{aligned} \tag{7}$$

otherwise $M(x_i, \phi_{k-\{c+c_n\}}) = 0$. By combining with Eq.2 in the main paper, we have

$$\begin{aligned} y_i = c \leftrightarrow \widetilde{y}_i = c \wedge (\forall c_n \in D_{c-ns}, p(c|x_i) > p(c_n|x_i)) \\ \leftrightarrow y_i = c \wedge (\forall c_n \in D_{c-ns}, M(x_i, \phi_{k-\{c+c_n\}}) = 1). \end{aligned} \tag{8}$$

1.3 SFT^{+k}

SFT [14] detects noisy samples according to predictions stored in a memory bank $\mathcal{M}$. $\mathcal{M}$ contains the last T epochs' predictions of each sample. A sample x_i is detected as noisy if a fluctuation event occurs, *i.e.*, the sample classified correctly at the previous epoch t_1 is misclassified at t_2, where $t_1 < t_2$. The occurrence of the fluctuation event can be formulated as $fluctuation(x_i, y_i) = 1$, otherwise $fluctuation(x_i, y_i) = 0$ *i.e.*,

$$\begin{aligned} & \text{fluctuation}(x_i, y_i) = 1 \\ \leftrightarrow\ & \exists t_1, t_2 \in \{t - T, \cdots, T\} \wedge t_1 < t_2 \\ & s.t.\ f_\theta(x_i)^{t_1} = \widetilde{y}_i \wedge f_\theta(x_i)^{t_2} \neq \widetilde{y}_i, \end{aligned} \tag{9}$$

where $f_\theta(x_i)^{t_1}$ represents the prediction of x_i at epoch t_1. SFT is a probability-distribution-based approach and can fit our probabilistic model as follows. The propositional logic of SFT is,

$$p(c|x_i) = \begin{cases} 1 & : \quad \widetilde{y}_i = c \wedge \text{fluctuation}(x_i, \widetilde{y}_i) = 0 \\ 0 & : \quad otherwise. \end{cases} \tag{10}$$

I.e., SFT^{+k} applies the noise source distribution knowledge to SFT by stricting the constraints of fluctuation. The fluctuation events only occur when the previous correct prediction is misclassified as the noise source label. Thus, we define SFT^{+k} fluctuation as,

$$\begin{aligned} &\text{fluctuation}(x_i, y_i, D_{y_i - ns}) = 1 \\ \leftrightarrow\ &\exists c_n \in D_{y_i - ns}, \exists t_1, t_2 \in \{t - T, \cdots, T\} \wedge t_1 < t_2, \\ &s.t.\ f_\theta(x_i)^{t_1} = y_i \wedge f_\theta(x_i)^{t_2} = c_n. \end{aligned} \tag{11}$$

Combining Eq.2 in the main paper, Eq. 10 and Eq. 11, SFT^{+k} detects x_i with clean label $y_i = \widetilde{y}_i = c$ with $p(c|x_i)$ as:

$$\begin{aligned} & y_i = c \\ \leftrightarrow\ & \widetilde{y}_i = c \wedge p(c|x_i) > \max(\{p(c_n|x_i)|c_n \in D_{c-ns}\}) \\ \leftrightarrow\ & \widetilde{y}_i = c \wedge p(c|x_i) = 1 \\ \leftrightarrow\ & \widetilde{y}_i = c \wedge fluctuation(x_i, \widetilde{y}_i, D_{\widetilde{y}_i - ns}) = 0. \end{aligned} \tag{12}$$

1.4 UNICON^{+k}

UNICON [5] estimate the clean probability by using Jensen-Shannon divergence (JSD) d_i, which is a measure of distribution disagreement. JSD is defined by KLD, which is the Kullback-Leibler divergence function. We follow the same JSD definition as UNICON in the adaptation method. Given the predicted probability p_i and label $\widetilde{y}_i$, $d_i = JSD(\widetilde{y}_i, p_i)$. The value of d_i ranges from 0 to 1 and the smaller the d_i is, the higher the probability of $\widetilde{y}_i$ being clean. A cutoff value d_{cutoff} is used to select clean samples. To summarize, the propositional logic of UNICON is,

$$\begin{aligned} & p(c|x_i) = 1 - JSD(x_i, y_i) \\ \leftrightarrow\ & y_i = c \wedge JSD(x_i, y_i) < d_{cutoff} \end{aligned} \tag{13}$$

otherwise $p(c|x_i) = 0$. Then noise source knowledge is integrated with our unified framework:

$$\begin{aligned} y_i = c \leftrightarrow\ & \widetilde{y}_i = c \wedge p(c|x_i) > \max(\{p(c_n|x_i)|c_n \in D_{c-ns}\}) \\ \leftrightarrow\ & \widetilde{y}_i = c \wedge (\forall c_n \in D_{c-ns}, JSD(x_i, \widetilde{y}_i) < JSD(x_i, c_n)). \end{aligned} \tag{14}$$

1.5 DISC^{+k}

DISC [8] employs weak and strong augmentations on each single noisy labeled data and divides samples into *Clean*, *Hard*, and *Purified* sets according to the prediction confidences on the two-augmentation views. The clean set is determined with the prediction confidence in weak view $Conf_w$, confidence in strong view $Conf_s$, and dynamic instance specific thresholds ($DIST$) for weak and strong views τ_w and τ_s. The $DIST$ is defined as,

$$\tau(x, t) = \lambda\tau(t-1) + (1-\lambda)\max(\{Conf(c, x)|c \in K\}, \tau(0) = 0. \tag{15}$$

where $Conf(c, x)$ represents the prediction confidence in sample x for class c. Then the clean set selected at time t of DISC can be defined as,

$$C(t) = \{x_i | Conf_w(\widetilde{y}_i, x_i) > \tau_w(x_i, t)\} \cap \{x_i | Conf_s(\widetilde{y}_i, x_i) > \tau_s(x_i, t)\}. \quad (16)$$

According to the Eq.1 in the main paper, sample x_i is clean at time t in DISC is detected with,

$$y_i = c \leftrightarrow \widetilde{y}_i = c \wedge (Conf_w(\widetilde{y}_i, x_i) > \tau_w(x_i, t) \wedge Conf_s(\widetilde{y}_i, x_i) > \tau_s(x_i, t)). \quad (17)$$

DISC^{+k} introduces class-wise comparisons and makes adaptations to the instance-specific threshold. Instead of taking the maximum confidence from all the class (*i.e.* $\max(\{Conf(c, x)|c \in K\})$, DISC^{+k} only selects maximum from the labeled class and noise source classes. To be specific, we have,

$$\tau^{+k}(x, \widetilde{y}, t) = \lambda\tau(t-1) + (1-\lambda)\max(\{Conf(c, x)|c \in D_{\widetilde{y}-ns}\} \cup \{Conf(\widetilde{y}, x)\}). \quad (18)$$

According to the Eq.2 in the main paper, the noise source knowledge is integrated as,

$$y_i = c \leftrightarrow \widetilde{y}_i = c \wedge (Conf_w(\widetilde{y}_i, x_i) > \tau_w^{+k}(x_i, \widetilde{y}_i, t) \wedge Conf_s(\widetilde{y}_i, x_i) > \tau_s^{+k}(x_i, \widetilde{y}_i, t)). \quad (19)$$

2 Datasets

2.1 CIFAR datasets [7] with synthesized noise

Dominant noise There are *recessive* and *dominant* classes in dominant noise. For CIFAR-10 [7], category index of the last 5 are *recessive* classes and the first five are *dominant* classes. In other words, category index 6-10 samples might be mislabeled as label index 1-5. Different numbers of samples are mixed for different noise ratios so that the dataset is still balanced after mislabeling. Table 1 shows the number of samples per category for each noise ratio. Notably, the dataset is balanced after the mislabeling. In each recessive class, there are multiple noise sources, with all dominant classes serving as the noise sources. To illustrate, in CIFAR-10 [7], classes 6-10 are considered recessive, and instances of these classes might be incorrectly labeled as dominant classes 1-5. To maintain balance after mislabeling, we adopt an unbalanced sampling approach to construct the dataset. For instance, with a noise ratio of 0.5 in the CIFAR-10 dataset, we sample 1250 instances for each *dominant* class and 3750 instances for each *recessive* class. After mislabeling 1250 samples to the *dominant* class for each *recessive* class, there are 2500 samples in each class.

Asymmetric noise. Labels are corrupted to visually similar classes. Pair (C_1, C_2) represents the samples in class C_1 and C_2 are possibly mislabeled as each other. Noise ratios in the experiments are only the noise ratio in class pairs, *i.e.* not the overall noise ratio. Here are the class pairs of CIFAR-10 and CIFAR-100 [7] for asymmetric noise. **CIFAR-10** [7] (trucks, automobiles), (cat, dog), (horse, deer). **CIFAR-100** [7] (beaver, otter), (aquarium fish, flatfish), (poppies, roses), (bottles, cans), (apples, pears), (chair, couch), (bee, beetle), (lion, tiger), (crab, spider), (rabbit, squirrel), (maple, oak), (bicycle, motorcycle).

Table 1: Sample composition for CIFAR-10/CIFAR-100 [7] dominant noise.

CIFAR-10 [7] Dominant Noise			
Noise ratio	0.2	0.5	0.8
Dominant class	2000	1250	500
Recessive class	3000	3750	4500
CIFAR-100 [7] Dominant Noise			
Noise ratio	0.2	0.5	0.8
Dominant class	200	125	50
Recessive class	300	375	450

2.2 Cell dataset BBBC036 [1]

For our experiments, we subsampled 100 treatments to evaluate natural noise. Table 2 shows the treatment list. ("NA" refers to the *control* group, *i.e.* no treatment group.)

Table 2: Treatments used from the BBBC036 dataset [1]

NA	BRD-K88090157	BRD-K38436528	BRD-K07691486	BRD-K97530723
BRD-A32505112	BRD-K21853356	BRD-K96809896	BRD-A82590476	BRD-A95939040
BRD-A53952395	BRD-A64125466	BRD-A99177642	BRD-K90574421	BRD-K07507905
BRD-K62221994	BRD-K62810658	BRD-K47150025	BRD-K17705806	BRD-K85015012
BRD-K37865504	BRD-A52660433	BRD-K66898851	BRD-K15025317	BRD-K37392901
BRD-K91370081	BRD-K39484304	BRD-K03842655	BRD-K76840893	BRD-K62289640
BRD-K14618467	BRD-K52313696	BRD-K43744935	BRD-K86727142	BRD-K21680192
BRD-K06426971	BRD-K24132293	BRD-K68143200	BRD-K08554278	BRD-K78122587
BRD-A47513740	BRD-K18619710	BRD-A67552019	BRD-K17140735	BRD-K30867024
BRD-K36007650	BRD-K51318897	BRD-K90382497	BRD-K00259736	BRD-K95435023
BRD-K52075040	BRD-K03642198	BRD-K47278471	BRD-K17896185	BRD-K95603879
BRD-A70649075	BRD-K02407574	BRD-A90462498	BRD-K67860401	BRD-A64485570
BRD-K88429204	BRD-A49046702	BRD-K50841342	BRD-K35960502	BRD-K77171813
BRD-K54095730	BRD-K93754473	BRD-K22134346	BRD-K72703948	BRD-K31342827
BRD-K31542390	BRD-K18250272	BRD-K00141480	BRD-K37991163	BRD-K13533483
BRD-K67439147	BRD-A91008255	BRD-K39187410	BRD-K26997899	BRD-K89732114
BRD-K50135270	BRD-K95237249	BRD-K44849676	BRD-K20742498	BRD-K31912990
BRD-K96799727	BRD-K09255212	BRD-A89947015	BRD-K78364995	BRD-K49294207
BRD-K08316444	BRD-K89930444	BRD-K50398167	BRD-K47936004	BRD-A72711497
BRD-A97104540	BRD-A50737080	BRD-K80970344	BRD-K50464341	BRD-K97399794

2.3 Cell dataset CHAMMI-CP [2]

Three compounds with a *control* group are selected for our experiments: BRD-A29260609 (weak reaction), BRD-K04185004 (medium reaction), and BRD-K21680192 (strong reaction).

2.4 Clothing1M dataset [15]

We conducted experiments on the Clothing1M dataset [4], the noise source knowledge is summarized according to the confusion matrix from the dataset [4]. We use $a \rightarrow b$ to represent a as the noise source of b. The prior noise knowledge is: *Chiffon* $\rightarrow$ *Shirt*, *Sweater* $\rightarrow$ *Knitwear*, *Knitwear* $\rightarrow$ *Sweater*, *Jacket* $\rightarrow$ *Windbreaker*, *Windbreaker* $\rightarrow$ *Down coat*, and *Vest*$\rightarrow$ *Dress*.

3 Feature extractors for each dataset

We used a pre-trained ResNet34 [3] on CIFAR-10/CIFAR-100 [7] for all approaches (UNICON [5] trains on two networks), ResNet50 [3] on Animal-10N [12] and Clothing1M [15] datasets. For experiments on BBBC036 [1] we used an Efficient B0 [13] for all methods and all methods used ConvNet [10] for CHAMMI-CP [2] dataset. To support the 5 channel images in cell datasets, we replaced the first convolutional layer in the network to support the new image dimensions.

4 Hyperparameters

For a fair comparison, we use the same hyperparameter settings as in prior work [5, 6, 8, 11, 14] for CIFAR-10/CIFAR-100 [7] datasets. Hyperparameters of the cell dataset BBBC036 [1] were set via grid search using the validation set. All the experiments use the same batch size of 128. "fl-ratio" of CRUST [11] and CRUST^{+k}, which controls the size of selected clean samples is set as the same as the noise ratio in synthesized noise and set as 0.6 in cell dataset BBBC036 [1] and CHAMMI-CP [2], 0.9 in Animal10N [12] and Clothing1M [15]. All the other hyperparameters for each dataset are summarized in Table 3.

Table 3: Hyperparameters for each dataset.

	learning rate	warm-up epochs	total number of epochs
CIFAR-10/CIFAR-100 [7]	1e-2	40	120
BBBC036 [1]	2e-4	10	100
CHAMMI-CP [2]	2e-4	5	30
Animal10N [12]	5e-3	3	30
Clothing1M [15]	5e-2	0	200

5 Additional results on lower noise ratio of dominant noise

We also performed experiments with 0.2 dominant noise on CIFAR-10/CIFAR-100 [7] datasets. The results in Table 4 demonstrate that knowledge integration is also beneficial in cases of lower noise ratios, showcasing the broad applicability of LNL+K across a range of noise levels from 0.2 to 0.8.

Table 4: 0.2 Dominant noise results on CIFAR-10 and CIFAR-100 dataset. The best test accuracy is marked in bold, and the better result between LNL and LNL+K methods is marked with underlined. We find incorporating source knowledge helps in almost all cases.

	CIFAR-10 [7]	CIFAR-100 [7]
Baseline	85.47±0.52	50.37±0.45
DualT [16]	86.55±0.06	34.88±0.11
GT-T	88.09±0.04	59.32±0.14
SOP [9]	89.86±0.40	62.47±0.47
CRUST [11]	88.21±0.22	53.48±0.80
CRUST^{+k}	89.53±0.05	58.69±0.50
FINE [6]	86.23±0.30	53.68±1.54
FINE^{+k}	88.69±0.06	57.22±1.16
SFT [14]	89.48±0.21	51.82±0.67
SFT^{+k}	89.78±0.03	54.36±0.48
UNICON [5]	90.82±0.14	63.28±0.32
UNICON^{+k}	90.83±0.11	66.77±0.54
DISC [8]	93.10±0.12	69.75±0.13
DISC^{+k}	**93.55**±0.03	**70.02**±0.30

6 Ethical considerations

This study was conducted with biological images of human bone osteosarcoma cells, an immortalized cell line used for research purposes only. The images or data in this study do not contain patient information of any kind. The use of these images, and the algorithms to analyze them, is to test the effects of treatments. Automating drug discovery has positive impacts on society, specifically the potential to help find cures for diseases of pressing need around the world in shorter times, and utilizing fewer resources. The proposed methods could be used to optimize drugs that harm people; we do not intend that as an application, and we expect regulations in biological labs to prevent such uses.

References

1. 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)
2. 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)

3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Kaneko, T., Ushiku, Y., Harada, T.: Label-noise robust generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2467–2476 (2019)
5. 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)
6. 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)
7. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
8. 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)
9. 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)
10. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
11. 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)
12. 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)
13. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
14. 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)
15. 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)
16. 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)