Appendix

A. Pseudocode of the Proposed Method

Algorithm 2 details the training procedure of the proposed framework for label-noise learning with intrinsically long-tailed data.

```
Algorithm 2: The training process of the proposed learning framework
Input: Noisy training data $\mathcal{D} = \{x_i, \hat{y}_i\}_{i=1}^N$
Initialize the model $\theta$ trained on $\mathcal{D}$;
while $e <$ MaxIterationNumber do
    for $c = 1$ to $M$ do
        for $i = 1$ to $|\mathcal{D}_c|$ do
            Obtain and store the feature $\mathbf{f}_i$ and the prediction confidence $\mathbf{p}_i$ for $x_i$ by model $\theta$;
        end
        Calculate the average prediction confidence $\bar{\mathbf{p}}_c$ for class $c$;
    end
    for $c = 1$ to $M$ do
        Calculate the confidence thresholds $H_c$ for class $c$ by Equation (6);
        Calculate the adaptive centroid $\mathbf{o}_c$ for class $c$ by Equation (5);
        for $i = 1$ to $|\mathcal{D}_c|$ do
            Calculate the additional weight $W(x_i)$ by Equation (2);
            Calculate $WJSD(x_i)$ by Equation (3);
            Calculate $ACD(x_i)$ by Equation (9);
        end
        Apply GMM with values of bi-dimensional metrics to all the samples in class $c$;
        Adopt dimension selection for class $c$ by Algorithnm 1;
        Adopt cluster selection for selected dimension by Equation (10) to generate two sets $\mathcal{D}_c^{clean}$ and $\mathcal{D}_c^{noisy}$;
    end
    Update the model $\theta$ by SSL training with $\mathcal{D}^{clean}$ as labeled data and $\mathcal{D}^{noisy}$ as unlabeled data.
end
```

B. Proof of Theorem 1

Proof. Suppose x_i and x_j are two samples in class c, p_i^c and q_i^c are the c's dimension of their prediction confidence $\mathbf{p}_i = [p_i^1, p_i^2, ..., p_i^M]$ and $\mathbf{p}_j = [p_j^1, p_j^2, ..., p_j^M]$, respectively. Their common observed class label $\hat{\mathbf{y}} = [\hat{y}_j^1, \hat{y}_j^2, ..., \hat{y}_j^M]$ is in the one-hot form where only the value of the c's dimension is 1 ($\hat{y}_i^c = 1$), and the values on other dimensions are all 0 ($\hat{y}_i^d = 0$ for all $d \neq c$). Jensen-Shannon Divergence (JSD) for sample x_i is defined as:

$$
\begin{aligned}
JSD(x_i) =& \frac{1}{2} KL\left(\mathbf{p}_i \middle\| \frac{\mathbf{p}_i + \hat{\mathbf{y}}_i}{2}\right) + \frac{1}{2} KL\left(\hat{\mathbf{y}}_i \middle\| \frac{\mathbf{p}_i + \hat{\mathbf{y}}_i}{2}\right) && (11) \\
=& \frac{1}{2}\left(\sum_{d=1}^{M} p_i^d \log \frac{2p_i^d}{p_i^d + \hat{y}_i^d} + \sum_{d=1}^{M} \hat{y}_i^d \log \frac{2\hat{y}_i^d}{p_i^d + \hat{y}_i^d}\right) && (12) \\
=& \frac{1}{2}\left(\sum_{d \neq c} p_i^d \log \frac{2p_i^d}{p_i^d} + p_i^c \log \frac{2p_i^c}{p_i^c + \hat{y}_i^c} + \hat{y}_i^c \log \frac{2\hat{y}_i^c}{p_i^c + \hat{y}_i^c}\right) && (13) \\
=& \frac{1}{2}\left(\sum_{d \neq c} p_i^d + p_i^c \log \frac{2p_i^c}{p_i^c + \hat{y}_i^c} + \log \frac{2}{p_i^c + 1}\right) && (14) \\
& && (15)
\end{aligned}
$$

$$
\begin{aligned}
&=\frac{1}{2}\left(1-p_i^c+p_i^c\log\frac{2p_i^c}{p_i^c+1}+\log\frac{2}{p_i^c+1}\right) &(16)\\
&=\frac{1}{2}\left(1-p_i^c+p_i^c+p_i^c\log p_i^c-p_i^c\log\left(p_i^c+1\right)+1-\log\left(p_i^c+1\right)\right) &(17)\\
&=\frac{1}{2}\left(2+p_i^c\log p_i^c-\left(p_i^c+1\right)\log\left(p_i^c+1\right)\right). &(18)
\end{aligned}
$$

The base of logarithm is 2 for the above derivation. Then

$$
\begin{aligned}
|JSD(x_i)-JSD(x_j)| &= \left|\frac{1}{2}\left(2+p_i^c\log p_i^c-\left(p_i^c+1\right)\log\left(p_i^c+1\right)\right)-\frac{1}{2}\left(2+p_j^c\log p_j^c-\left(p_j^c+1\right)\log\left(p_j^c+1\right)\right)\right| &(19)\\
&= \frac{1}{2}\left|\left(p_i^c\log p_i^c-\left(p_i^c+1\right)\log\left(p_i^c+1\right)\right)-\left(p_j^c\log p_j^c-\left(p_j^c+1\right)\log\left(p_j^c+1\right)\right|\right. &(20)
\end{aligned}
$$

Now, for $u\in(0,1)$, let

$$
f(u)=u\log u-(u+1)\log(u+1). \tag{21}
$$

The first- and second-order derivative of $f(u)$ can be obtained by:

$$
\begin{aligned}
f'(u) &= \left(\log u+u\cdot\frac{1}{u}\cdot\frac{1}{\ln 2}\right)-\left(\log(u+1)+(u+1)\cdot\frac{1}{u+1}\cdot\frac{1}{\ln 2}\right) &(22)\\
&= \log u-\log(u+1). &(23)\\
f''(u) &= \left(\frac{1}{u}-\frac{1}{u+1}\right)\frac{1}{\ln 2}. &(24)
\end{aligned}
$$

It can be observed that $f'(u)<0$ and $f''(u)>0$ that makes $f'(u)$ monotonically increase and $|f'(u)|$ monotonically decrease. By Lagrange mean value theorem, if function $f(u)$ is continuous and differentiable on the interval $(0,1)$, then there is at least one point ξ between two real numbers $a,b\in(0,1)$:

$$
|f(a)-f(b)|=|f'(\xi)(a-b)|=|f'(\xi)|\cdot|a-b|. \tag{25}
$$

As $\xi>a$ and $|f'(u)|$ monotonically decrease, we have:

$$
\begin{aligned}
|f'(\xi)|\cdot|a-b| \le& |f'(a)|\cdot|a-b| &(26)\\
=& |\log a-\log(a+1)|\cdot|a-b| &(27)\\
=& \left|\log\frac{a}{a+1}\right|\cdot|a-b| &(28)\\
=& \log\frac{a+1}{a}\cdot|a-b| &(29)
\end{aligned}
$$

By replacing $f(u)$ by $JSD(x)$, a by p_i^c and b by p_j^c, we have:

$$
|JSD(x_i)-JSD(x_j)|\le\frac{1}{2}\log\left(\frac{p_i^c+1}{p_i^c}\right)|p_i^c-p_j^c|. \tag{30}
$$

C. The Correlation between Purity and Clean Ratio

To validate the promotion effect of purity on the separation of clean and noisy samples, we show the highest proportion of clean samples for clusters obtained with different purity in the tail class with asymmetric noise. Fig. A1 describes the relationship between purity and the proportion of clean samples in the cluster. It can be observed that there is a positive correlation between purity and the proportion of clean samples in the cluster. Therefore, enhancing the purity of centroid can effectively improve the separation ability of clean and noisy samples.

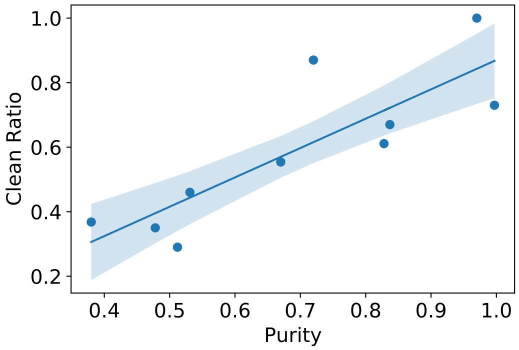


Figure A1: The correlation between purity and clean ratio.

D. The Purity by High-Confidence Sample Set

In order to validate the effectiveness of our proposed high-confidence sample set in terms of purity improvement, we compared the purity difference before and after the selection of high-confidence samples in CIFAR-10 with 0.4 asymmetric noise and 0.01 imbalance factor. As shown in Fig. A2, the purity of all classes is significantly improved through the selection of high-confidence samples, and the improvement of the tail class is more prominent. It illustrates the robustness of our approach to the long-tail distribution.

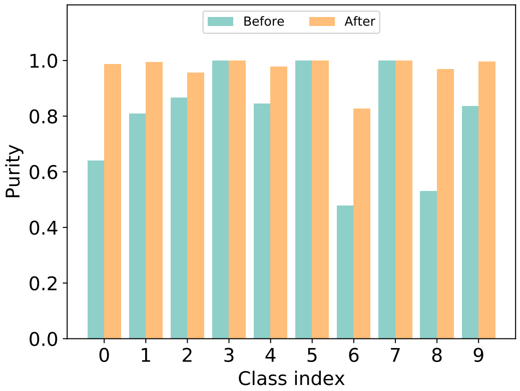


Figure A2: The purity difference before and after the selection of high confidence samples.

Dataset	CIFAR-10		CIFAR-100	
Noise Ratio	0.4			
Imbalance Factor	0.1	0.01	0.1	0.01
baseline-clean	87.88	70.11	59.75	42.65
cRT-full	84.62	74.53	55.94	46.22

Table A1: Performance comparison with asymmetric noise and long-tail distribution.

E. The Effectiveness of Feature in Asymmetric Noise

In order to show why the feature is good enough to discriminate the sample under asymmetric noise, we retrain the classifier using full clean samples (cRT-full) after fixing the features for the model trained under asymmetric noise with long-tail distribution. Accordingly, we also use clean and long-tail distribution samples to train the model as a baseline for

Dataset		CIFAR-10		CIFAR-100	
Imbalance Factor		0.01			
Noise Ratio (**Sym.**)		0.4	0.6	0.4	0.6
Baseline	CE	47.81	28.04	21.99	15.51
LT	LA	42.63	36.37	21.54	13.14
	LDAM	45.52	35.29	18.81	12.65
	IB	49.07	32.54	20.34	12.10
NL	DivideMix	32.42	34.73	36.20	**26.29**
	UNICON	61.23	54.69	32.09	24.82
NL-LT	MW-Net	46.62	39.33	19.65	13.72
	RoLT	60.11	44.23	23.51	16.61
	HAR	51.54	38.28	20.21	14.89
	ULC	45.22	50.56	33.41	25.69
Our	TABASCO	**62.34**	**55.76**	**36.91**	26.25

Dataset		CIFAR-10		CIFAR-100	
Imbalance Factor		0.01			
Noise Ratio (**Asym.**)		0.2	0.4	0.2	0.4
Baseline	CE	56.56	44.64	25.35	17.89
LT	LA	58.78	43.37	32.16	22.67
	LDAM	61.25	40.85	29.22	18.65
	IB	56.28	42.96	31.15	23.40
NL	DivideMix	41.12	42.79	38.46	29.69
	UNICON	53.53	34.05	34.14	30.72
NL-LT	MW-Net	62.19	45.21	27.56	20.04
	RoLT	54.81	50.26	32.96	-
	HAR	62.42	51.97	27.90	20.03
	ULC	41.14	22.73	34.07	25.04
Our	TABASCO	**62.98**	**54.04**	**40.35**	**33.15**

Table A2: Performance comparison with synthetic noise and long-tail distribution. The best results are shown in bold.

comparison. Tab. A1 reports the accuracy of both approaches. It can be observed that the potential of the model under the influence of noise and long-tail distribution is similar to the performance of the model trained directly under the clean and long-tail distribution samples. It means that the model can still learn good features to distinguish between samples even under the dual effects of the noise and long-tail distribution. It is also an important guarantee to use features to separate samples under long-tail distribution and asymmetric noise.

F. More Experimental Results on CIFAR-10/100

In this section, we compare the performance of different methods with symmetric and asymmetric noise in 0.01 imbalance factor, which is a more difficult scenario. Tab. A2 reports the accuracy of different methods in these settings. One of the results of RoLT [49] is empty because the code does not run properly in this case. It can be observed that our proposed method achieves the best performance in most cases, which further verifies the effectiveness of our method.

Dataset		Red		10N	100N
Imbalance Factor		≈ 0.01		0.01	
Noise Ratio		0.2	0.4	≈ 0.4	
Baseline	CE	30.88	31.46	49.31	25.28
LT	LA	10.32	9.560	50.09	26.39
	LDAM	14.30	15.64	50.36	30.17
	IB	16.72	16.34	56.41	31.55
NL	DivideMix	33.00	34.72	30.67	31.34
	UNICON	31.86	31.12	59.47	37.06
NL-LT	MW-Net	30.74	31.12	54.95	31.80
	RoLT	15.78	16.90	61.23	33.48
	HAR	32.60	31.30	56.84	32.34
	ULC	34.24	34.84	43.89	35.71
Our	TABASCO	**37.20**	**37.12**	**64.54**	**39.30**

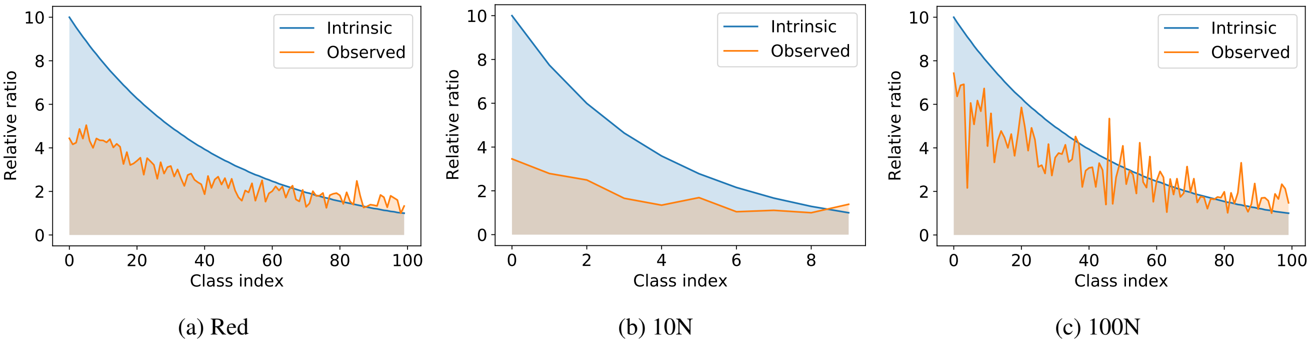
Table A3: Performance comparison with real-world noise and long-tail distribution. The best results are shown in bold.

G. More Experimental Results on Benchmarks

In this section, we compare the performance of different methods with realist noise in 0.01 imbalance factor, which is a more difficult scenario. Tab. A3 reports the accuracy of different methods in these settings. It can be observed that our proposed method achieves the best performance in all cases, which further verifies the effectiveness of our method.

H. The Observed and Intrinsic Distribution of Benchmarks

In this section, we plot the observed and intrinsic distribution of benchmarks we proposed in Fig. A3. We use benchmarks with an imbalance factor of approximately 0.1 and a noise ratio of approximately 0.4 for our analysis. Relative ratio is denoted by the ratio between the number of samples in each class and the minimum number of samples. It can be observed that the intrinsic and observed distributions of all benchmarks are significantly different, and observed distribution is more balanced. This means that it is difficult to focus on the intrinsic tail class and the tail class has a high percentage of noise, making it more difficult to separate noise and clean samples in the intrinsic tail class. It also corresponds to the two key challenges of the problem we present and demonstrates that our proposed benchmarks are a good measure of the effectiveness of different approaches to the problem.

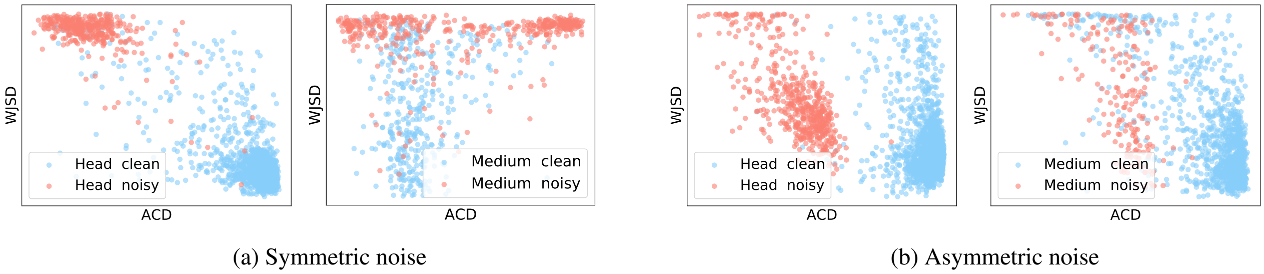


(a) Red (b) 10N (c) 100N

Figure A3: The observed and intrinsic distribution of benchmarks.

I. More Sample Distribution on Bi-dimensional Separation Metric

In this section, we plot the values of bi-dimensional metrics of both clean and noisy samples for head and medium classes under different noise types in Fig. A4. For the case of symmetric noise shown in Fig. A4 (a), it can be observed that both WJSD and ACD can well separate clean and noisy samples in the head class, while ACD cannot effectively distinguish clean samples from noisy samples in the medium class. In this case, WJSD shows its advantage in sample separation. For the case of asymmetric noise shown in Fig. A4 (b), it can be observed that WJSD cannot distinguish clean samples from noisy samples well for head and medium classes, while ACD can distinguish them well. It further validates the complementarity of the bi-dimensional metrics.



(a) Symmetric noise (b) Asymmetric noise

Figure A4: Scatter plot of the values of the proposed bi-dimensional metrics with (a) symmetric noise and (b) asymmetric noise.

J. Hyperparameter Sensitivity Analysis

We investigate the impact of η for dimension selection. The results are shown in Fig. A5. We vary η from 0.4 to 0.8, and the relative accuracy varies between -0.5% and 0.5%. When η is relatively large, more cases not suitable for WJSD separation are WJSD separated, thus compromising the sample separation effect and leading to a decrease in model performance. Similarly, when η is relatively small, cases suitable for WJSD separation are filtered out, affecting model performance. In general, there is little variation in the accuracy of the model, so the dimension selection we proposed is robust for η.

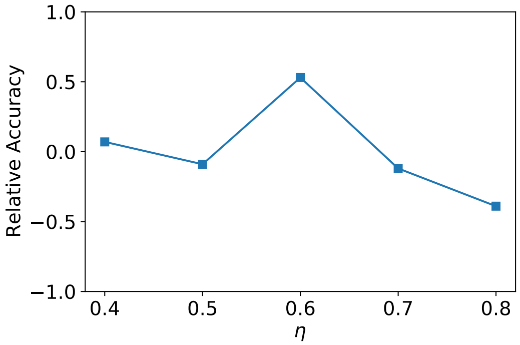


Figure A5: Impact of hyperparameter η.

K. Computational Cost Analysis

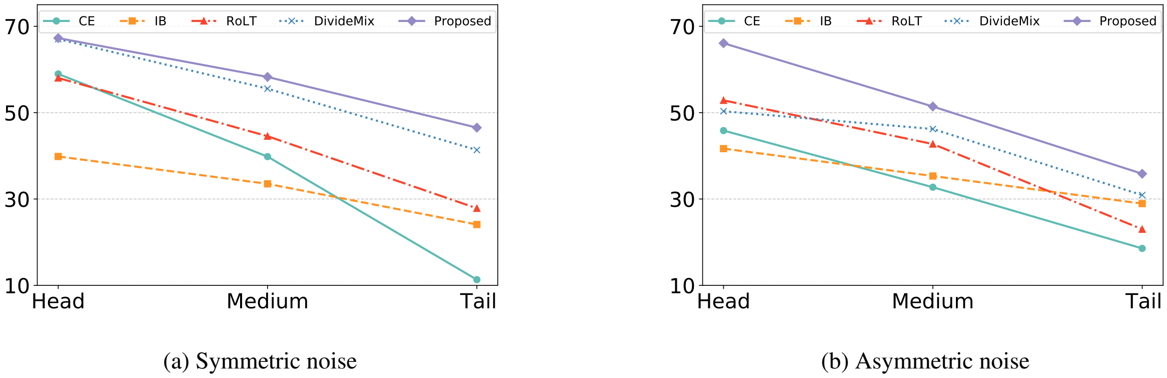
We compute the relative training cost of each method compared to the baseline on CIFAR-100, as presented in the Tab. A4. By comparison, it is evident that the training cost of the proposed method is indeed higher but falls within an acceptable range. Despite this higher cost, the proposed method offers remarkable performance gains in return.

Method	MW-Net	RoLT	HAR	DivideMix	UNICON	Proposed
Relative Cost	×2.60	×2.06	×1.48	×2.28	×2.54	×4.49

Table A4: Relative training cost compared to the baseline.

L. Experimental Results across Classes

We conducted a performance comparison of different methods on CIFAR-100 with an imbalance factor of 0.1 and a noise ratio of 0.4 under head, medium and tail classes in Fig. A6. It can be seen that our approach effectively improves the performance of tail classes without sacrificing the performance of head classes. The underlying rationale behind is that the proposed WJSD can effectively separate tail classes samples without affecting the separation of head classes, and ACD are not solely focused on improving the tail classes.



(a) Symmetric noise

(b) Asymmetric noise

Figure A6: Line plot of the performance across classes with (a) symmetric noise and (b) asymmetric noise.