Supplementary Material
NGC: A Unified Framework for Learning with Open-World Noisy Data

Zhi-Fan Wu[1,2]*, Tong Wei[1]*, Jianwen Jiang[2]*, Chaojie Mao[2], Mingqian Tang[2], Yu-Feng Li[1]†
[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]Alibaba Group, China
{wuzf, weit}@lamda.nju.edu.cn, liyf@nju.edu.cn
{jianwen.jjw, chaojie.mcj, mingqian.tmq}@alibaba-inc.com

A. Experimental Details

In this section, we introduce the experiment details. We first introduce the out-of-distribution (OOD) datasets used in our experiments. Then, we present the experimental settings of our method. Finally, we provide details about the evaluation metrics used for evaluating the classification and OOD detection performance of our method.

A.1. Out-of-Distribution Datasets

We use the OOD datasets below in our experiments:

- **TinyImageNet**. The Tiny ImageNet dataset contains 50,000 training images from 200 different classes, which are drawn from the original 1,000 classes of ImageNet. We randomly choose samples from training set and resize each image to 32 × 32.
- **Places-365**. The Places-365 dataset has 365 scene categories and there are 900 images per category in the test set. The OOD samples are randomly chosen from test set of Places-365 and resize to 32 × 32.

A.2. Experimental Setup

For all CIFAR experiments, we train PreAct ResNet-18 network for 300 epochs using SGD with the momentum 0.9 and weight decay $5 \cdot 10^{-4}$. The initial learning rate is set to 0.15 and cosine decay schedule is used. The batch size is set to 512. The dimension of projector layer is set to 64. The temperature parameter is fixed as $\tau_1 = 0.3$ and $\tau_2 = 1.0$. For CIFAR-10 experiments, we use $k = 30$ for sym. noise and $k = 10$ for asym. noise, warmup with cross-entropy loss without other components for 5 epoch. For all CIFAR-100 experiments, we use $k = 200$, warmup for 30 epoch for CIFAR-100 datasets. For parameter η, in LOND task, we use 0.8 for all experiments, and in closed-world noisy label task, we set it to 0.7 for CIFAR-10 and 0.6 for CIFAR-100.

For Webvision-50 dataset, most of hyperparameters are the same with CIFAR experiments except we set $k = 100$, $\eta = 0.8$. We train the inception-resnet v2 model using SGD following prior works. The initial learning rate is set to 0.2 and the batch size is 256. We train the network for 80 epochs and the warmup stage lasts 15 epochs.

A.3. Evaluation Metrics

We use the following three performance metrics to evaluate the performance.

- **Classification Accuracy.** The top-1 classification accuracy is calculated as the mean accuracy over all known (IND) classes. Predictions of data are obtained as the classes with the highest softmax probabilities.
- **AUROC.** AUROC is the Area Under the Receiver Operating Characteristic curve and can be calculated by the area under the TPR against FPR curve.
- **F-measure.** The F-measure (F) is calculated as 2 times the product of precision (p) and recall (r) divided by the sum of p and r:

$$F = 2 \cdot \frac{p \cdot r}{p + r}. \tag{1}$$

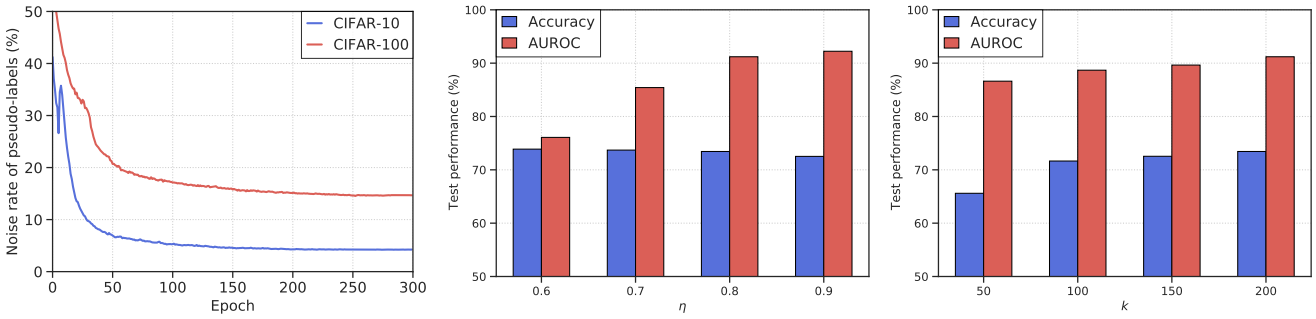
p is calculated as true positive over the sum of T_p and false positive:

$$p = \frac{T_p}{T_p + F_p}. \tag{2}$$

r is calculated as T_p over the sum of T_p and false negative:

$$r = \frac{T_p}{T_p + F_n}. \tag{3}$$

*Equal contribution. †Corresponding author. This work was supported by Alibaba Group through Alibaba Innovative Research Program and the National Natural Science Foundation of China (61772262).



(a) Effectiveness of noise correction. (b) Impact of hyperparameter η (c) Impact of hyperparameter k

Figure 1: Experimental results. (a) Effectiveness of noise correction. Both CIFAR-10 and CIFAR-100 datasets are under 50% sym. noise. (b-c) Analysis of the impact of hyperparameters under 50% IND noise (CIFAR-100), 20k and 10k OOD noise (Places-365) in training set and test set, respectively. η is for confidence-based selection and k is for k-NN graph.

B. Additional Experimental Results

In this section, we first show the visualization results of feature representation and subgraph selection, which demonstrate the validity of our methods. Then we present the effectiveness of graph-based noise correction. We also analyze the sensitivity of hyperparameters. In addition, the performance of model ensemble and the impact of AugMix on WebVision-50 is provided. Finally, we compare NGC with recent related work, ProtoMix [1] on LOND task.

B.1. Visualization Results

Visualization of learned representation. We visualize the learned feature representations of our method and DivideMix via t-SNE in Figure 2. CIFAR-10 with 50% sym. noise is used as IND dataset and 20k OOD samples are added in each experiment. We use CIFAR-100, TinyImageNet, and Places-365 as OOD datasets for each experiment, respectively. The points in brown represent OOD samples, while samples with other colors are from CIFAR-10. Figures 2a to 2c show the learned representations of DivideMix, which are extracted from the last layer of the model. For comparison, Figures 2d to 2f visualize the output of the projector Proj in our method. It can be observed that our method can learn more meaningful representations and separate OOD samples from IND samples effectively.

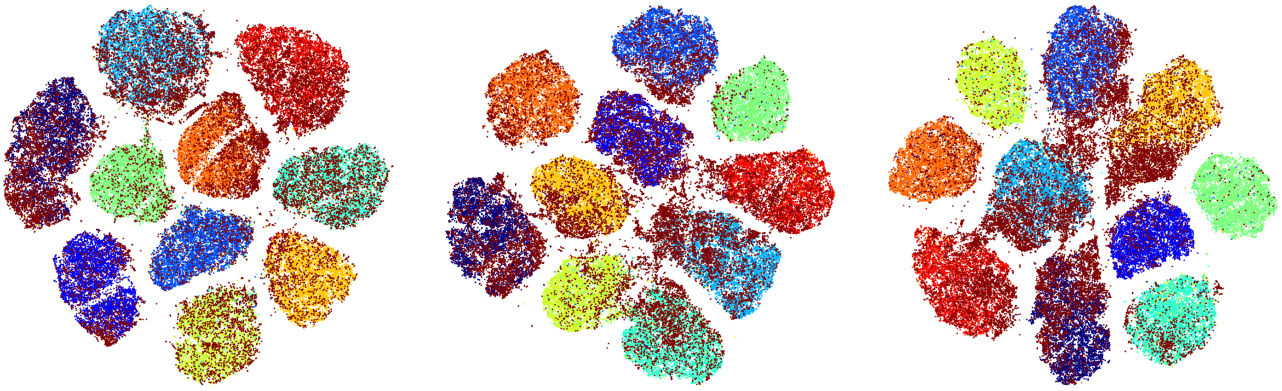
Visualization of subgraph selection. To further justify the efficacy of the proposed subgraph selection, we visualize the k-NN graph obtained at different training iterations in Figure 3. CIFAR-10 with 50% sym. noise is used as IND dataset and 20k CIFAR-100 data are added as OOD samples. We draw all the samples with pseudo-label 1. In these graphs, we use green points to represent samples removed by confidence-based selection while black points are samples removed by geometry-based selection. Points in yellow represent clean data selected by our method. The edges included in the largest connected component are in red. At different training iterations, we visualize the constructed k-NN graph (top row) and the refined graph (bottom row) by performing our confidence-based selection. As the training progresses, the feature representations of IND and OOD samples are gradually separated. Moreover, it can be seen that confidence-based selection significantly degrades the connectivity between clean samples and OOD samples, which can be further beneficial to geometry-based selection. As a consequence, samples retained by geometry-based selection distribute more and more compact in feature space. This observation justifies the validity of subgraph selection.

B.2. Effectiveness of Noise Correction

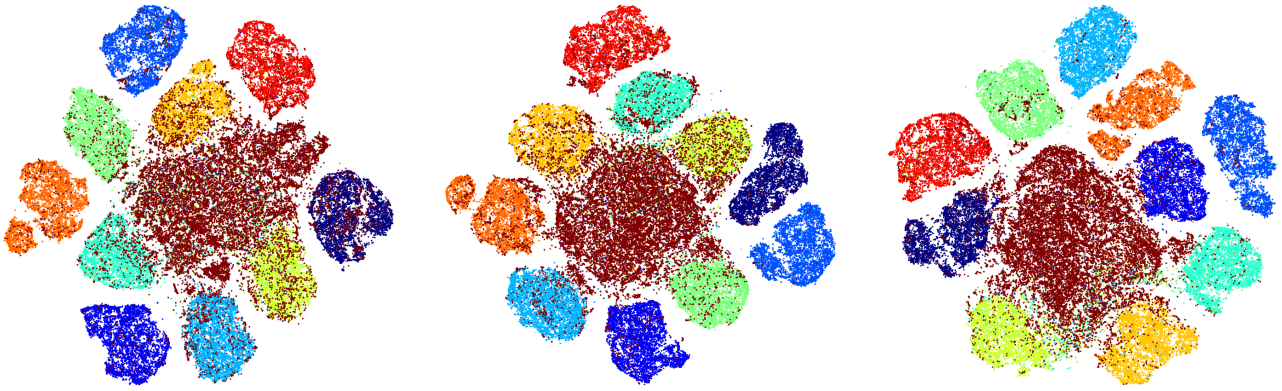
We demonstrate the effectiveness of graph-based noise correction on CIFAR-10 and CIFAR-100 datasets with 50% symmetric noise. As shown in Figure 1a, As the training progresses, the noise rate continues decreasing. Our method reduces noise rate from 50% to 4.24% for CIFAR-10 and 14.67% for CIFAR-100. This validates our noise correction methods can correct noisy labels effectively.

B.3. Hyperparameter Sensitivity Analysis

Analysis of η and k. We investigate the impact of η for confidence-based selection and k which is used to construct the k-NN graph. The results are shown in Figure 1b and 1c. We vary η from 0.6 to 0.9, and the test accuracy increases from 70% to 72%, showing that a small confidence threshold results in more label noise being included. AUROC increases from 72.08 to 92.23, this is because a higher threshold η can filter out more OOD noisy samples, which can be further beneficial for representation learning and the calculation of prototypes. As for the parameter k, we choose its value from $\{50, 100, 150, 200\}$. It can be seen that NGC achieves similar performance with differ-



(a) CIFAR-100 as OOD (DivideMix) (b) TinyImageNet as OOD (DivideMix) (c) Places-365 as OOD (DivideMix)



(d) CIFAR-100 as OOD (Ours) (e) TinyImageNet as OOD (Ours) (f) Places-365 as OOD (Ours)

Figure 2: t-SNE visualization of learned feature representation. CIFAR-10 with 50% sym. noise is used as IND dataset and 20k OOD samples are added for all experiments. The OOD samples are represented by brown points.

ent values except $k = 50$. The reason is that when k is too small, the k-NN graph is very sparse, resulting in fewer data points being obtained from the largest connected component, hence only a few clean samples are selected for training.

Analysis of ζ. We report F-measure under best threshold ζ in Table 1. Even with fixed ζ from 0.5 to 0.7, our method is robust enough and outperforms other methods with their best values of ζ in most cases. Here we report results for $\zeta = 0.5$ and $\zeta = 0.7$. We also report the standard deviation of best ζ, which shows the stability of our method.

B.4. Performance of Model Ensemble

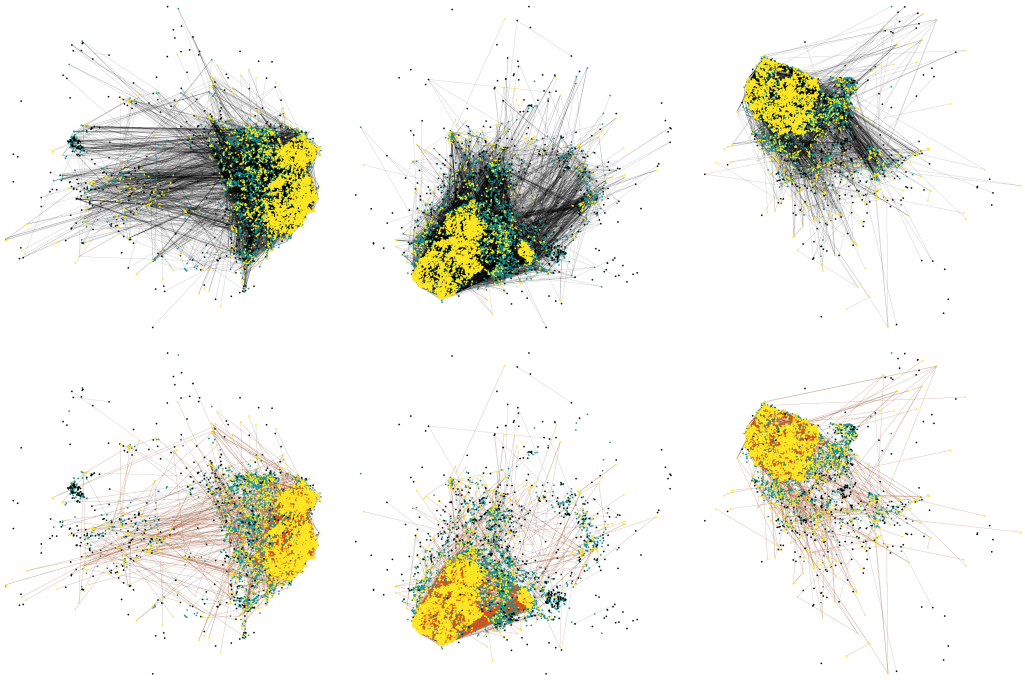
Since model ensemble has shown to be useful when dealing with noisy data, we ensemble the outputs of two networks during testing phase and report the results in Table 2. The complete DivideMix (DM) is used for comparison. Results show that our method outperforms DivideMix in most cases.

Table 1: F-measure (threshold ζ). IND dataset is with 50% symmetric noise, 20k and 10k OOD samples are added into training set and test set, respectively. **Bold**: best; Underlined: 2nd & 3rd.

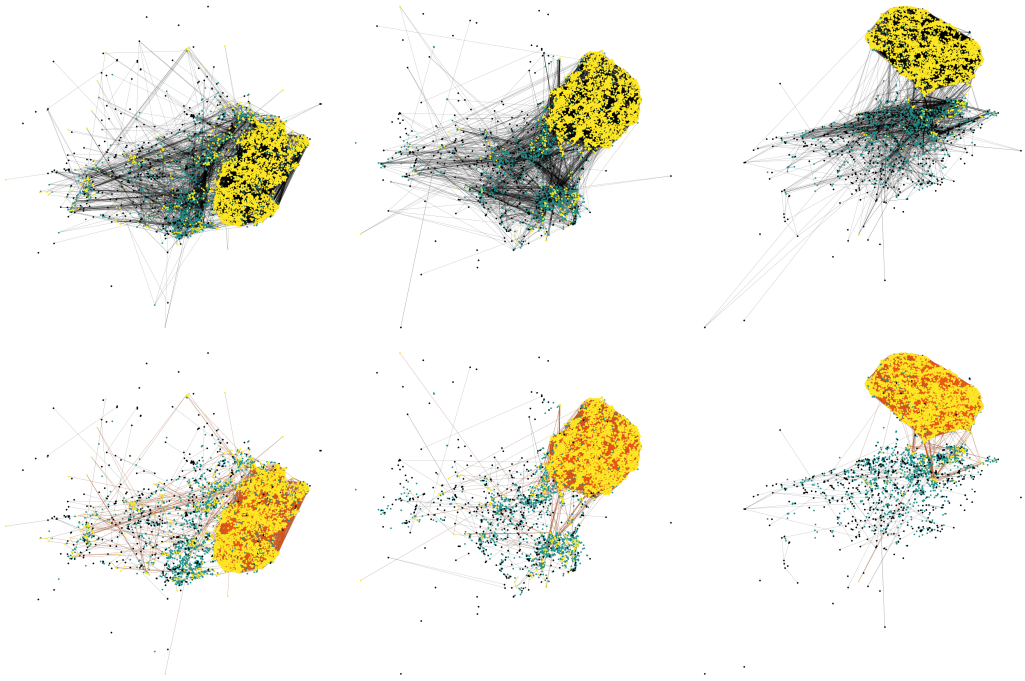
IND	OOD	MPS	ODIN	MD	Ours	Ours$_{\zeta=0.50}$	Ours$_{\zeta=0.70}$
C-10	C-100	0.698$_{(0.81)}$	0.681$_{(0.83)}$	0.635$_{(0.36)}$	**0.838$_{(0.55)}$**	0.835$_{(0.50)}$	0.788$_{(0.70)}$
	TIN	0.726$_{(0.83)}$	0.707$_{(0.85)}$	0.702$_{(0.33)}$	**0.875$_{(0.54)}$**	0.873$_{(0.50)}$	0.802$_{(0.70)}$
	P-365	0.717$_{(0.81)}$	0.705$_{(0.14)}$	0.651$_{(0.40)}$	**0.887$_{(0.56)}$**	0.882$_{(0.50)}$	0.827$_{(0.70)}$
C-100	TIN	0.687$_{(0.41)}$	0.705$_{(0.02)}$	0.526$_{(0.38)}$	**0.773$_{(0.67)}$**	0.743$_{(0.50)}$	0.770$_{(0.70)}$
	P-365	0.685$_{(0.37)}$	0.696$_{(0.01)}$	0.541$_{(0.39)}$	**0.731$_{(0.70)}$**	0.687$_{(0.50)}$	0.731$_{(0.70)}$
ζ (stand. dev.)		0.21	0.39	0.02	0.07	0.00	0.00

B.5. Impact of AugMix on WebVision-50

To better reveal the superiority of our method, we conduct ablation studies for AugMix on WebVision-50 dataset. The results are reported in Table 3. First, it can be seen that



(a) 50 epoch (b) 100 epoch (c) 150 epoch



(d) 200 epoch (e) 250 epoch (f) 300 epoch

Figure 3: t-SNE visualization of the proposed subgraph selection at different training iterations. CIFAR-10 with 50% sym. noise is used as IND dataset and 20k CIFAR-100 data are added as OOD samples. We draw all samples with pseudo-label 1. Green points represent samples removed by confidence-based selection and black points are samples removed by geometry-based selection. Points in yellow represent clean data selected by our method. Edges in the largest connected component are colored red. We visualize the constructed k-NN graph (top row) and the refined graph (bottom row) by performing our confidence-based selection.

AugMix does help enhance the performance. Second, without applying AugMix, our method consistently outperforms strong baselines, i.e., ELR and DivideMix. The results further demonstrate the effectiveness of our method.

Table 2: Test accuracy (%) using model ensemble. $^+$ indicates ensemble models.

Data	CIFAR-10					CIFAR-100			
Type	Sym.				Asym.	Sym.			
Ratio	20%	50%	80%	90%	40%	20%	50%	80%	90%
DM	95.0	93.7	**92.4**	74.2	**91.4**	74.8	72.1	57.6	29.2
Ours	**95.88**	**94.54**	91.59	**80.46**	90.55	**78.98**	**75.91**	**62.70**	**29.76**
DM$^+$	95.7	94.4	**92.9**	75.4	**92.1**	76.9	74.2	59.6	31.0
Ours$^+$	**96.27**	**95.09**	92.20	**83.75**	91.70	**81.08**	**77.16**	**64.00**	**34.18**

Table 3: Ablation study for AugMix on WebVision-50. $^+$ indicates ensemble models.

Method	WebVision		ILSVRC12	
	top-1	top-5	top-1	top-5
Ours (w/ AugMix)	79.16	91.84	74.44	91.04
ELR	76.26	91.26	68.71	87.84
Ours (w/o AugMix)	**77.56**	**91.36**	**72.92**	**91.32**
DivideMix$^+$	77.32	91.64	**75.20**	90.84
ELR$^+$	77.78	91.68	70.29	89.76
Ours$^+$ (w/o AugMix)	**79.08**	**91.80**	75.12	**91.72**

B.6. Comparison with ProtoMix

As one of the most recent related works, ProtoMix [1] employs unsupervised contrastive loss and mixup prototypical contrastive loss to learn robust representations, which can address different types of noisy data. We report the comparison results of NGC and ProtoMix on LOND task in Table 4. For all experiments, we inject 50% symmetric IND noise. 20k and 10k OOD samples are randomly selected and added into training set and test set, respectively. Although ProtoMix is not designed to detect OOD examples at test time, it is natural to achieve this by measuring the similarity between test examples and class prototypes, as shown in Eq. (9) in the main text. From the results, we can observe that NGC achieves better or comparable results in test accuracy. Regarding AUROC and F-measure, NGC consistently outperforms ProtoMix in all cases. Recall that, ProtoMix identifies IND and OOD noise according to predictive confidence, which means samples with high predictive confidence are determined as clean. As a result, many noisy samples are likely to be misidentified as DNNs gradually fit the training data. NGC overcomes this problem by exploiting the geometric structure of data. For each class, confident samples that clustered together are further selected by calculating the largest connected component. Our belief is that clean samples of the same class should distribute closely to each other, while noisy samples are pushed away. By first performing confidence-based selection, it breaks the connection between noisy and clean samples in the graph, which facilitates our geometry-based selection. Consequently, NGC excludes more noisy samples from training and achieves better performance.

Table 4: Performance comparison of ProtoMix and NGC (Ours) on LOND task. 50% symmetric IND noise is injected into training set, 20k and 10k OOD samples are added into training set and test set, respectively.

IND	OOD	Accuracy	AUROC	F-measure
		ProtoMix / NGC		
C-10	C-100	**92.51** / 92.31	84.64 / **90.37**	0.783 / **0.838**
	TIN	93.12 / **93.54**	93.47 / **94.18**	0.862 / **0.875**
	P-365	92.76 / **93.67**	94.14 / **94.31**	0.868 / **0.887**
C-100	TIN	72.80 / **73.49**	78.58 / **94.24**	0.653 / **0.773**
	P-365	72.05 / **73.44**	75.19 / **91.20**	0.624 / **0.731**

C. Pseudo-code of Our Proposed Method

Algorithm 1 lists the pseudo-code of NGC. For a better understanding of the proposed method, we illustrate the whole process in Figure 4.

References

[1] Junnan Li, Caiming Xiong, and Steven Hoi. Learning from noisy data with robust representation learning, 2021.

Algorithm 1 Noisy Graph Cleaning Procedure (one epoch)

Input: training dataset $\{(\boldsymbol{x}_i, y_i)_{i=1}^N\}$, k-NN parameter k, confidence threshold η.
Construct the k-NN graph G on training samples.
Refine soft pseudo-label $\tilde{\boldsymbol{Y}}_i$ for each sample $\boldsymbol{x}_i$ by performing graph-based noise correction on G.
If $\max_k \tilde{\boldsymbol{Y}}_{ik} < \eta$ and $\tilde{\boldsymbol{Y}}_{iy_i} \leq \frac{1}{K}$, remove the point $\boldsymbol{x}_i$ and its adjacent edges from the graph.
The resulting graph is denoted by $\tilde{G}$.
Initialize the set of clean data $S = \emptyset$.
for $k = 1 \cdots K$ **do**
    Remove points that do not belong to class k from graph $\tilde{G}$, i.e., $\hat{y}_i \neq k, \forall i \in [N]$.
    The resulting graph is denoted by $\tilde{G}(k)$.
    Determine the connected components of $\tilde{G}(k)$ by disjoint-set data structures.
    Remove small connected components of the graph $\tilde{G}(k)$, that is, only the largest connected component is retained.
    The resulting graph is denoted by $\tilde{G}(k)_{lcc}$. Points in $\tilde{G}(k)_{lcc}$ are treated as clean samples.
    Update clean data set $S = S \cup \tilde{G}(k)_{lcc}$.
end for
Calculate cross-entropy loss and subgraph-level contrastive loss on S.

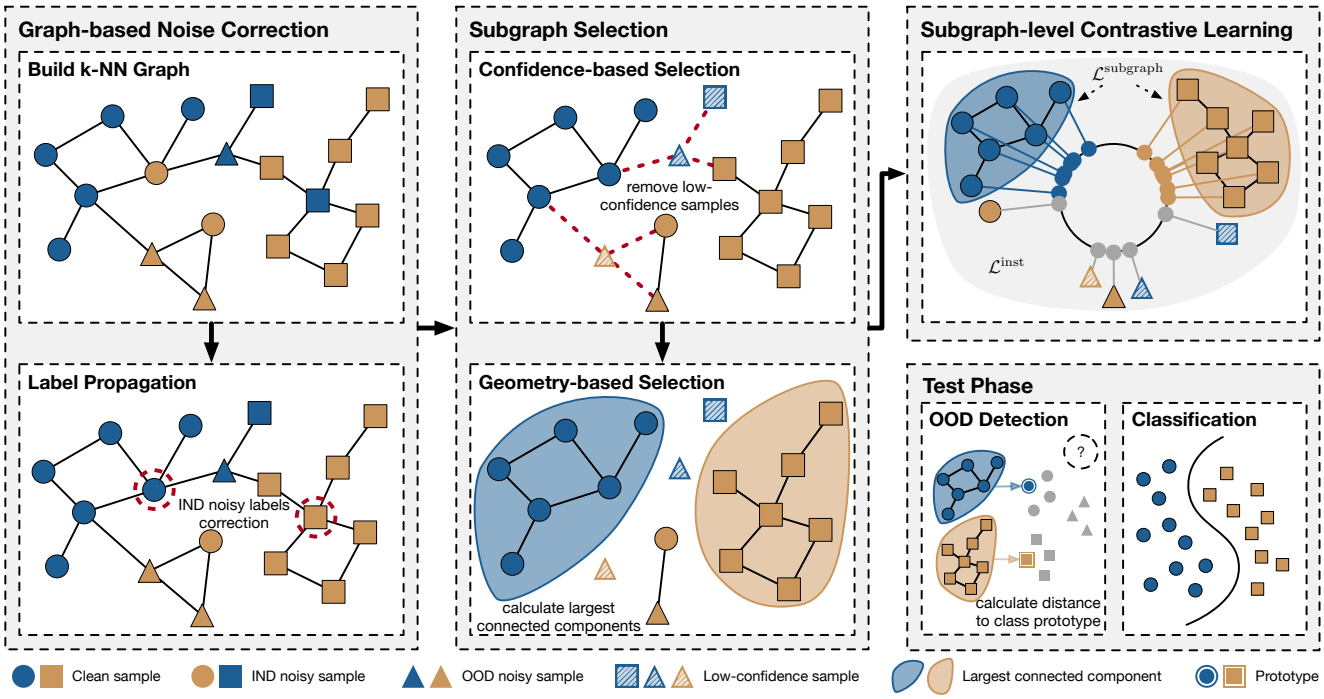


Figure 4: An illustration of proposed framework in binary classification case.