# COMPLEMENTARY LABEL LEARNING WITH POSITIVE LABEL GUESSING AND NEGATIVE LABEL ENHANCEMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Complementary label learning (CLL) is a weakly supervised learning paradigm that constructs a multi-class classifier only with complementary labels, specifying classes that the instance does not belong to. We reformulate CLL as an inverse problem that infers the full label information from the output space information. To be specific, we propose to split the inverse problem into two subtasks: *positive label guessing* (PLG) and *negative label enhancement* (NLE), collectively called PLNL. Specifically, we use well-designed criteria for evaluating the confidence of the model output, accordingly divide the training instances into three categories: highly-confident, moderately-confident and under-confident. For highly-confident instances, we perform PLG to assign them pseudo labels for supervised training. For moderately-confident and under-confident instances, we perform NLE by enhancing their negative label set with different levels and train them with the augmented negative labels iteratively. In addition, we unify PLG and NLE into a consistent framework, in which we can view all the pseudo-labeling-based methods from the perspective of negative label recovery. We prove that the error rates of both PLG and NLE are upper bounded, and based on that we can construct a classifier consistent with that learned by clean full labels. Extensive experiments demonstrate the superiority of PLNL over the state-of-the-art CLL methods, e.g., on STL-10, we increase the classification accuracy from 34.96% to 55.25%. The code has been submitted to supplementary material.

## 1 INTRODUCTION

Over the past few years, large-scale and accurately labeled data has tremendously boosted the development of deep neural networks. However, collecting accurately labeled data is extremely time-consuming, labor-intensive and sometimes requires specific expertise in real-world tasks. To reduce the dependency on large-scale and accurately labeled datasets, deep learning communities have given increasing attention to weakly supervised learning, including but not limited to partial label learning (Cour et al., 2011; Xie and Huang, 2018; Feng and An, 2019; Lv et al., 2020; Xia et al., 2023; Huang and Cheung, 2024; He et al., 2024; Tian et al., 2024), noisy label learning (Natarajan et al., 2013; Han et al., 2018; Song et al., 2022; Wei et al., 2020; Zhang et al., 2023; Huang et al., 2023), semi-supervised learning (Van Engelen and Hoos, 2020; Sohn et al., 2020; Xie et al., 2020; Yang et al., 2022; Li et al., 2023b; Xie et al., 2023), positive-unlabeled learning (Niu et al., 2016; Kiryo et al., 2017).

Here, we consider a recently proposed weakly supervised learning framework called complementary label learning (CLL) (Ishida et al., 2017; Feng et al., 2020). In CLL, each training instance is associated with one or multiple complementary labels (CLs) which specify one or multiple classes that the instance does not belong to. The goal of CLL is to learn a multi-class classifier only from complementary labeled data. In real-world scenario, if the number of classes is huge, choosing the correct class label from many candidate classes is difficult and laborious, while choosing one or several of the incorrect class labels as CLs would be much easier and thus less costly. Recently, CLL has been applied to online learning (Kaneko et al., 2019), medical image segmentation (Rezaei et al., 2020) and medical molecular imaging (Tapper et al., 2024), etc. Besides, another promising future application scenario of CLL is to ensure privacy security in data collection scenarios. For example,

collecting some survey data may require extremely private questions and it would be mentally less demanding if we ask the respondent to provide some incorrect answers as CLs (Dwork, 2008).

Previous studies on CLL can be roughly divided into two categories: methods that attempt to construct an unbiased risk estimator (URE-based) (Ishida et al., 2017; 2019; Feng et al., 2020) and methods based on feature learning (FL-based) (Chou et al., 2020; Wang et al., 2021; Liu et al., 2022; Jiang et al., 2024). For URE-based methods, Ishida *et al.* (Ishida et al., 2017) and Feng *et al.* (Feng et al., 2020) showed that the ordinary classification risk can be recovered by their proposed unbiased risk estimator only from complementary labeled data. Ishida *et al.* (Ishida et al., 2019) later extended the unbiased risk estimator to arbitrary losses and models. Chou *et al.* (Chou et al., 2020) proposed a surrogate complementary loss framework, which avoids the extremely noisy gradient problem encountered in unbiased risk estimator. For FL-based methods, Wang *et al.* (Wang et al., 2021) gave the first attempt to leverage regularization techniques with complementary label by aligning the model output of one instance and its multiple augmented views. Liu *et al.* (Liu et al., 2022) proposed to integrate self-supervised and self-distillation to complementary learning. Jiang *et al.* (Jiang et al., 2024) leveraged a contrastive learning framework to facilitate CLL. These methods mainly focus on the design of robust loss functions or the exploration of feature space information, while neglecting the power of output space information.

We propose that CLL can be viewed as solving the multi-class classification problem from two inverse aspects, where one is to infer the positive label and another is to infer the negative labels. To this end, we propose two subtasks: *positive label guessing* (PLG) and *negative label enhancement* (NLE). We use well-designed criteria for evaluating the confidence of the model output, accordingly divide the training instances into three categories: highly-confident, moderately-confident and under-confident in each epoch. We perform PLG by simply pseudo-labeling highly-confident instances for supervised training. Unlike pseudo-labeling methods used in semi-supervised learning (SSL), PLG pseudo-labeling reaches high selected ratio and high precision even without any positive labels available.

More importantly, previous SSL methods lack the utilization of untrustworthy instances. They either discard this part or simply employ techniques such as consistency regularization. In this paper, we perform NLE by enhancing the negative label set of moderately-confident and under-confident instances and train them with the augmented negative labels iteratively.

Although PLG and NLE will inevitably bring pseudo-labeling errors, we theoretically prove that the error rates are upper bounded. And the generalization error of the learned classifier under PLG and NLE errors is also upper bounded, which means that we can construct a classifier consistent with that learned by clean full labels. We demonstrate that PLNL achieves state-of-the-art performance on five benchmark datasets. Our contributions can be summarized as follows:

- *A novel method for CLL*. Different from conventional loss design methods, we pioneer a novel method for CLL called PLNL that formulates CLL from output space information and solve it by two subtasks: PLG and NLE.
- *A unified framework for pseudo-labeling-based methods*. From the perspective of negative label recovery, we construct a unified framework for pseudo-labeling-based methods. We empirically show that PLNL outperforms state-of-the-art SSL methods in terms of pseudo-labeling error, selected ratio and recovered negative labels.
- *Solid theoretical analysis*. We theoretically prove that both the error rates of PLG and NLE are upper bounded. The generalization error of the learned classifier is also upper bounded.
- *State-of-the-art performance*. Extensive experiments on five benchmark datasets demonstrate the superiority of PLNL over the state-of-the-art CLL methods.

## 2 PRELIMINARIES

**Ordinary Multi-Class Classification.** Let $\mathcal{X} \in \mathbb{R}^d$ denote the feature space with $d$ dimensions and $\mathcal{Y} = \{1, 2, ..., K\}$ denote the label space with $K$ classes. The precisely labeled instance $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is sampled from an unknown probability distribution $p(\boldsymbol{x}, y)$. The goal of ordinary multi-class classification is to learn a parameterized function $f(x) : \mathbb{R}^d \to \mathbb{R}^K$ that minimizes the

classification risk:

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\mathcal{L}(f(\boldsymbol{x}), y), \tag{1}$$

where $\mathbb{E}_{p(\boldsymbol{x},y)}$ refers to the expectation across all possible samples drawn from the distribution $p(\boldsymbol{x}, y)$, $\mathcal{L} : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ is a multi-class classification loss function. In this paper, we consider a common case where the function $f$ is a deep neural network with the softmax output layer, where $f(\boldsymbol{x})$ is considered as the output prediction confidence of the model on each class. Since the probability distribution $p(\boldsymbol{x}, y)$ is unknown, we use the empirical risk $\widehat{R}(f)$ to approximate $R(f)$. Assuming a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ is independently drawn from distribution $p(\boldsymbol{x}, y)$, then we have

$$\widehat{R}(f) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f(\boldsymbol{x}), y_i). \tag{2}$$

**Complementary Label Learning.** Different from the ordinary multi-class classification, in CLL, let $\{(\boldsymbol{x}_i, \bar{Y}_i)\}_{i=1}^N$ be the complementary labeled dataset, where $N$ is the dataset size, $\bar{Y}_i$ indicates the complementary (negative) label set of $\boldsymbol{x}_i$. Each complementary labeled instance $(\boldsymbol{x}, \bar{Y}) \in \mathcal{X} \times \mathcal{Y}$ is sampled from an unknown probability distribution $\bar{p}(\boldsymbol{x}, \bar{y})$. Our goal is to learn a classifier that minimizes the classification risk Eq. (1) only from complementary labeled training instances. Then the empirical risk becomes:

$$\widehat{R}(f) = \frac{1}{N}\sum_{i=1}^{N}\bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i), \bar{Y}_i), \tag{3}$$

where $\bar{\mathcal{L}}_{CLL}$ is a specially designed loss function for learning from only complementary labeled data.

## 3 Proposed Method

The overall framework of PLNL is show in Fig.1, and the pseudo-code is presented in Appendix A. We begin by employing weak and strong augmentation to a complementary labeled image $\boldsymbol{x}_i$, which leads to two augmented views $\boldsymbol{x}_i^w$, $\boldsymbol{x}_i^s$. These two images are then fed into a two-view network with shared weight $f(\boldsymbol{x}; \Theta)$ to obtain two prediction confidences $f(\boldsymbol{x}_i^w)$ and $f(\boldsymbol{x}_i^s)$. Then we utilize the two-view prediction confidences to select three subsets of training instances mentioned above, i.e., highly-confident, moderately-confident and under-confident. We select these subsets using the historical confidences of the previous training epochs to better alleviate confirmation bias. Finally, different techniques are utilized to conquer individual subsets. In this section, we first explain the well-designed confidence-based instances selection strategy, and then introduce the PLG for the highly-confident instances and two different versions of NLE for the moderately-confident instances and the under-confident instances in detail.

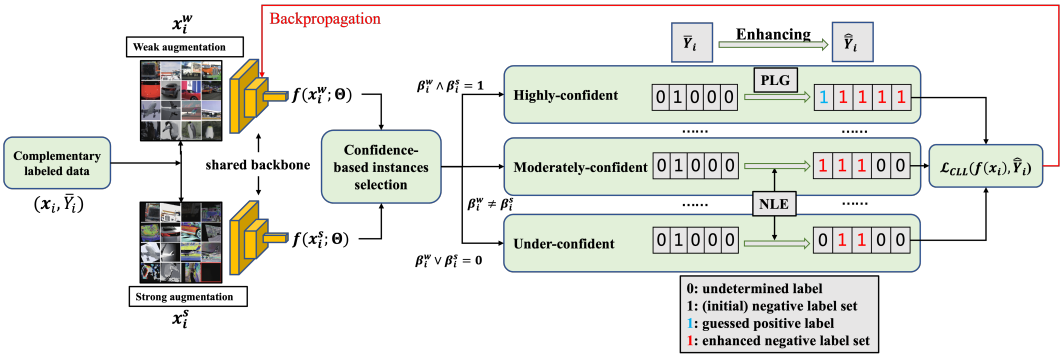



Figure 1: The overall framework of PLNL. We employs a two-view network (shared backbone) to extract features and compute confidences for weak and strong augmentations of one instance respectively. After the selection of highly-confident, moderately-confident and under-confident set, We employ PLG on highly-confident instances and NLE for the rest. The loss is computed on the enhanced labels of both views and the model updates through backpropagation.

### 3.1 CONFIDENCE-BASED INSTANCES SELECTION

We first maintain two memory banks $\boldsymbol{M}^w$ and $\boldsymbol{M}^s$ for weak and strong augmentation respectively, each with size $t \times N \times K$ to store the historical prediction confidence over the past $t$ epochs. For simplification, we only consider one view here unless otherwise specified.

$$\boldsymbol{M}_i = [f^1(\boldsymbol{x}_i), \ldots, f^t(\boldsymbol{x}_i)]. \tag{4}$$

where $p^t(\boldsymbol{x}_i)$ denotes the prediction confidence of the $t$-th epoch in the memory bank $\boldsymbol{M}$.

We propose to select subsets based on the following criteria.

$$\omega_{i1} = \forall 1 \leq j \leq t, \arg\max(f^j(\boldsymbol{x}_i)) \notin \bar{Y}_i, \tag{5}$$

$$\omega_{i2} = \forall 1 \leq j, k \leq t, \arg\max(f^j(\boldsymbol{x}_i)) = \arg\max(f^k(\boldsymbol{x}_i)), \tag{6}$$

$$\omega_{i3} = \forall 1 \leq j \leq t, \max(f^j(\boldsymbol{x}_i)) \geq \lambda, \tag{7}$$

where $\omega_{i1}, \omega_{i2}, \omega_{i3}$ are boolean variables which indicate whether the corresponding criterion is satisfied. $\omega_{i1}$ ensures that the label corresponding to the max prediction confidence does not fall on the complementary label set, which excludes the complementary labels from being selected as positive label. $\omega_{i2}$ ensures that the max prediction confidence be stable and show no sign of fluctuations over the past $t$ epochs. $\omega_{i3}$ ensures the max prediction confidence should be higher than a threshold $\lambda$. Note that $\lambda$ can be either a man-made threshold or a self-adaptive one, which will be discussed in detail later.

**Warm up.** Before selecting, we warm up the model using the entire training set. The goal of this stage is to reduce the classification risk and obtain some historical prediction confidence to construct the memory bank since we have no historical information at initial epoch. In this paper, we use SCL-LOG algorithm (Chou et al., 2020) to warm up models for 20 epochs.

**Instance-aware self-adaptive threshold.** The threshold in criterion $\omega_{i3}$, as is mentioned above, can be a fixed high threshold (like 0.95). However, a single global threshold does not consider the fitting difficulties of different instances, i.e., hard instances and easy instances. This will result in very few samples being selected in the early training stages as well as confirmation bias. Therefore, to comprehensively consider historical information, we design an instance-aware self-adaptive threshold for each instance at each epoch $t$ as:

$$\lambda(t) = \alpha\lambda(t-1) + (1-\alpha)f(t), \lambda(0) = \frac{1}{K}, \tag{8}$$

where $K$ is the number of classes, $f(t) = \max f(\boldsymbol{x})$ and $\alpha$ is the ratio which controls the threshold stability.

The threshold is initialized at a low value $\frac{1}{K}$, which will take more data into account and helps speed up convergence in the early stages. As the prediction confidence increases, the threshold grows higher to filter out wrong pseudo labels to alleviate the confirmation bias. Note that we compute $\lambda^w(t)$ and $\lambda^s(t)$ for two different views respectively according to Eq. (8). We use the momentum average confidence of each instance, computed based on all previous epochs. In this way, the threshold comprehensively considers historical information and remains stable and trustworthy.

**Subset Selection.** For the two-view network, we perform two independent verifications. Let $\beta_i^w = \omega_{i1} \wedge \omega_{i2} \wedge \omega_{i3}$ be the indicator of satisfying the criteria. Thus, $\beta_i^w$ and $\beta_i^s$ indicate whether the weak and strong views meet the criteria respectively. For an instance, if both views meet the criteria, we select it to the highly-confident subset, i.e.,

$$\mathcal{H} = \{\boldsymbol{x}_i | \beta_i^w \wedge \beta_i^s = 1\}, \tag{9}$$

It means the prediction confidences of both views are stable and high, thus we consider them to be highly-confident. The size of $\mathcal{H}$ is denoted as $N_h$.

Similarly, the moderately-confident subset consists of instances only one augmented version of which meet the criteria, i.e.,

$$\mathcal{M} = \{\boldsymbol{x}_i | \beta_i^w \neq \beta_i^s\}, \tag{10}$$

It means only one view's prediction is trustworthy, the other is not. which shows that the model are moderately-confident about its prediction. The size of $\mathcal{M}$ is denoted as $N_m$.

Finally, the under-confident subset consists of the rest of the instances, i.e.,

$$\mathcal{U} = \{\boldsymbol{x}_i | \beta_i^w \vee \beta_i^s = 0\}, \tag{11}$$

It means the prediction confidences of both views do not meet the designed criteria, these instances are considered under-confident. The size of $\mathcal{U}$ is denoted as $N_u$.

After obtaining $\mathcal{H}$, $\mathcal{M}$ and $\mathcal{U}$, we design different techniques to better utilize these different types of training instances.

### 3.2 Positive Label Guessing

For highly-confident set $\mathcal{H}$, we consider the label with the max prediction confidence as its positive label. **Conversely, all remaining labels are considered complementary labels.** Let $\hat{\bar{Y}}_i$ be the enhanced negative label set for instance $\boldsymbol{x}_i$, we have:

$$\hat{\bar{Y}}_i = \{c | c \in Y_i, c \neq \hat{y}_i\} \tag{12}$$

where $\hat{y}_i$ is the guessed positive label and $Y_i = \{1, 2, \ldots, K\}$ is the full label set.

For highly-confident set $\mathcal{H}$, we compute the CLL loss on the negative labels for both views:

$$\mathcal{L}_h = \frac{1}{N} \sum_{i=1}^{N_h} \bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i^w), \hat{\bar{Y}}_i) + \bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i^s), \hat{\bar{Y}}_i) \tag{13}$$

where $f(\boldsymbol{x}_i^w)$ and $f(\boldsymbol{x}_i^s)$ denote model outputs of weak augmentation and strong augmentation respectively.

### 3.3 Negative Label Enhancement

For the moderately-confident set $\mathcal{M}$ and the under-confident set $\mathcal{U}$, guessing the positive labels directly might lead to much more errors due to their relatively lower confidence. Therefore, we employ a different strategy for these instances, called negative label enhancement (NLE).

The rationale of NLE is that more negative labels will bring in additional supervision information for better training. However, whether the enhanced negative labels are correct remains a question. Intuitively, randomly enhancing negative labels will bring in a large number of labeling errors. To better enhance the reliability of NLE, we further exploit information in the output space and design the following solution.

**Calculation of $k$ Nearest Neighbor ($k$-NN) instances.** For instance $\boldsymbol{x}_i$ and its model output prediction confidence $y_i$, we can compute its $k$-NN instances in the output space. It is safe to assume that nearby instances in the output space should have the same positive label with a high probability, while their original complementary label sets are likely to vary. The formal definition of this assumption is as follows:

**Assumption 1.** *$\forall(\boldsymbol{x}_i, \bar{Y}_i) \in \mathcal{D}$ and its $k$-NN instances $(\boldsymbol{x}_i^{(j)}, \bar{Y}_i^{(j)})$, the positive label $y_i$ exists in its $k$-NN instances' complementary label set $\bar{Y}_i^{(j)}$ with probability no more than $\alpha_k$, any negative label $y_i' \neq y_i$ exist in its $k$-NN instances' complementary label set $\bar{Y}_i^{(j)}$ with probability no less than $\beta_k$.*

This assumption describes the intrinsic characteristics of CLL in the output space, which can be interpreted in two aspects. First, similarity in the input space will be mapped to similarity in the output space, which has been widely utilized for tackling representation learning problems (He et al., 2020). Second, instances of the same category are likely to be labeled with complementary labels of different categories, which is key to enhancing the negative labels.

**$k$-NN label frequency.** For instance $\boldsymbol{x}_i$, we propose to calculate the times a negative label appears in its $k$-NN instances' complementary label set and then enhance top-$\tau_i$ frequent ones, that is, add

them to the complementary label set of $\boldsymbol{x}_i$. We define the $j$-th $k$-NN label frequency of $\boldsymbol{x}_i$ as follows:

$$\boldsymbol{F}_{ij} = \sum\nolimits_{v=1}^{k} \mathbb{I}(j \in \bar{Y}_i^{(v)}), \tag{14}$$

where $\bar{Y}_i^{(v)}$ denotes the complementary label set of the $v$-th nearest instance of $\boldsymbol{x}_i$.

**Negative label enhancement.** We enhance the complementary label set $\bar{Y}_i$ by adding additional labels with top-$\tau_i$ label frequency. For each instance $\boldsymbol{x}_i$ in $\mathcal{M}$ and $\mathcal{U}$, the enhanced complementary (negative) label set $\widehat{\bar{Y}}_i$ is calculated by:

$$\widehat{\bar{Y}}_i = \{c | c \in \bar{Y}_i \vee c \in \text{top-}\tau_i\text{-max}_j(\boldsymbol{F}_{ij})\}. \tag{15}$$

However, the prediction confidence of the under-confident is more unreliable than that of the moderately-confident. Therefore, we should be more conservative when enhancing these instances as the $k$-NN information may be more unreliable. In our work, we set $\tau_i = \lceil \frac{K - s_i}{10} \rceil$ for $\mathcal{U}$ where $s_i$ is the size of $\bar{Y}_i$. For $\mathcal{M}$, we set $\tau_i = (1 + \frac{e}{E_{max}}) \lceil \frac{K - s_i}{10} \rceil$ where $e$ is current epoch, $E_{max}$ is total epochs. This provides a linear growing strategy for the moderately-confident because the model's output becomes increasingly accurate as the training progresses.

For moderately-confident set $\mathcal{M}$ and under-confident set $\mathcal{U}$, we compute the CLL loss on the negative labels for both views:

$$\mathcal{L}_{m,u} = \frac{1}{N} \sum_{i=1}^{N_m + N_u} \bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i^w), \widehat{\bar{Y}}_i) + \bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i^s), \widehat{\bar{Y}}_i) \tag{16}$$

where $f(\boldsymbol{x}_i^w)$ and $f(\boldsymbol{x}_i^s)$ denote model outputs of weak augmentation and strong augmentation respectively.

## 4 A Unified Framework for Pseudo-Labeling-Based Methods

Pseudo-labeling, which has been widely used in the recent semi-supervised learning (SSL) methods, is employed by giving unlabeled instances pseudo labels and train them in a supervised way. PLNL is an extension of pseudo-labeling. We not only consider pseudo-labeling of highly-confident instances, but also consider enhancing the negative label set of untrustworthy instances. In this way, we actually recover more supervised information than only leveraging pseudo-labeling and further boost the classification performance.

In this section, we construct a unified framework where PLG and NLE are viewed from the perspective of negative label recovery. Let $\hat{y}_i$ be the pseudo-label of $\boldsymbol{x}_i$. Let $Y_i$ be the full label space. Let $\widehat{\bar{Y}}_i$ be the reconstructed (PLNL) or imposed (pseudo-labeling) negative label set for $\boldsymbol{x}_i$.

For PLNL, PLG is equivalent to reconstructing a negative label set $\widehat{\bar{Y}}_i = \{c | c \in Y_i, c \neq \hat{y}_i\}$ of size $K - 1$, in which only the guessed positive label does not belong. NLE is equivalent to reconstructing a negative label set of size $s_i + \tau_i$, where we add $\tau_i$ negative labels to the original negative label set of size $s_i$.

Similarly, for pseudo-labeling methods, let the pseudo label for instance $\boldsymbol{x}_i$ be $\hat{y}_i$ as well. The process of pseudo-label highly-confident instances is also equivalent to imposing a negative label set $\widehat{\bar{Y}}_i = \{c | c \in Y_i, c \neq \hat{y}_i\}$ of size $K - 1$ as additional supervised information.

In this paper, we propose two metrics for evaluation of pseudo-labeling-based methods. Firstly, we define selected ratio $\eta$:

$$\eta = \frac{N_h}{N}, \tag{17}$$

Obviously, $\eta$ evaluates the ratio of highly-confident instances selected for pseudo-labeling methods.

Furthermore, we define average size of enhanced negative label set $\bar{s}$:

$$\bar{s} = \frac{\sum_{i=1}^{N} |\widehat{\bar{Y}}_i|}{N}, \tag{18}$$

In section 6, we empirically show that PLNL achieves lower error rate $\epsilon$, higher selection ratio $\eta$ and obviously larger size of negative label set $\bar{s}$ compared with pseudo-labeling method Fixmatch (Sohn et al., 2020) and Freematch (Wang et al., 2022).

# 5 Theoretical Analysis

## 5.1 Generalization Bound

For simplification, we only consider one view network here, which has no influence on the deduction of generalization error bound. Our goal is to learn a classification model $f(\boldsymbol{x};\Theta)$ by minimizing the empirical risk $\widehat{R}'(f)$ acquired from data with enhanced negative labels:

$$\widehat{R}'(f) = \frac{1}{N}\sum_{i=1}^{N}\bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i),\widehat{\bar{Y}}_i), \tag{19}$$

where $\widehat{\bar{Y}}_i$ denotes the enhanced negative label set of $\boldsymbol{x}_i$.

Let the CLL loss function be $\mathcal{L}_{CLL}(f(\boldsymbol{x}),\widehat{\bar{Y}}_i) = \sum_{y\notin\widehat{\bar{Y}}_i}(1/(K-|\widehat{\bar{Y}}_i|))\ell(f(\boldsymbol{x}),y)$ where $|\widehat{\bar{Y}}_i|$ is the size of the enhanced negative label set. Let $\bar{s} = \frac{\sum_{i=1}^{N}|\widehat{\bar{Y}}_i|}{N}$ be the average size of enhanced negative label set. Let $\epsilon_1 = \sum_{i=1}^{N_h}\frac{\mathbb{I}(y_i\in\widehat{\bar{Y}}_i)}{N_h}$ be the error rate of PLG. Let $\epsilon_2 = \sum_{i=1}^{N_m+N_u}\frac{\mathbb{I}(y_i\in\widehat{\bar{Y}}_i)}{N_m+N_u}$ be the error rate of NLE. The actual pseudo-labeling error rate $\epsilon = \frac{\sum_{i=1}^{N}\mathbb{I}(y_i\in\widehat{\bar{Y}}_i)}{N} = \frac{N_h}{N}\epsilon_1 + \frac{N_m+N_u}{N}\epsilon_2 = \eta\epsilon_1 + (1-\eta)\epsilon_2$. Moreover, $\ell(f(\boldsymbol{x}),y)$ is $\rho$-Lipschitz *w.r.t.* $f(\boldsymbol{x})$ where $\rho$ can be any Lipschitz constant (not necessarily the best). Let $\mathfrak{R}_N(\mathcal{F})$ be the expected Rademacher complexity (Mohri et al., 2018) of $\mathcal{F}$ with $N$ training instances. Let $\hat{f} = \text{argmin}_{f\in\mathcal{F}}\widehat{R}'(f)$ be the empirical risk minimizer, where $\mathcal{F}$ is a function class, and $f^* = \text{argmin}_{f\in\mathcal{F}}R(f)$ be the true risk minimizer. We derive the following theorem, which provides a generalization error bound for the proposed method.

**Theorem 1.** *Suppose that $\ell(f(\boldsymbol{x}),y)$ is bounded by $B$. For pseudo-labeling error rate $\epsilon\in(0,1)$, for any $\delta>0$, with probability at least $1-\delta$, we have*

$$R(\hat{f}) - R(f^*) \leq 2(1-\frac{1-\epsilon}{K-\bar{s}})B + 4\rho K\mathfrak{R}_N(\mathcal{F}) + 2KB\sqrt{\frac{log\frac{2}{\delta}}{2N}}, \tag{20}$$

***Remark.*** Detailed proofs are provided in Appendix B. Theorem 1 shows that as $N\to\infty$, $\epsilon_1\to 0$, $\epsilon_2\to 0$, the empirical risk minimizer converges to the true risk minimizer with high probability. It can be observed from Eq. (20) that the generalization bound is influenced by five factors: the number of categories $K$, the average size of enhanced negative label set $\bar{s}$ and two error rates. This is consistent with the intuition that more categories and less complementary labels will make the CLL problem harder. **In a nutshell, smaller PLNL pseudo-labeling error rates $\epsilon_1,\epsilon_2$ and larger size of enhanced negative label set $\bar{s}$ will produce better generalization performance.**

## 5.2 Error Bound of Positive Label Guessing

**Theorem 2.** *Suppose that $y_i$ denote the ground-truth positive label of $\boldsymbol{x}_i$ and $\hat{y}_i$ denote the guessed positive label which might not be true. PLG error rate $\epsilon_1$ is upper bounded by:*

$$\epsilon_1 = \mathbb{P}(y_i\in\widehat{\bar{Y}}_i) \leq (K-1-s_i)\psi, \tag{21}$$

where $K$ is class number, $s_i$ is the size of $\bar{Y}_i$ and $\psi\in(0,\frac{1}{K-1-s_i})$. Detailed proofs are provided in Appendix C.

## 5.3 Error Bound of Negative Label Enhancement

**Theorem 3.** *Suppose that $y$ denote the ground-truth positive label of $\boldsymbol{x}_i$ and $y'$ denote an arbitrary negative label. Let $\boldsymbol{F}_i^{(\tau_i)}$ denote the $\tau_i$-th largest label frequency. Let $p$ denote the probability of the*

*ground-truth positive label $y_i$ appearing in its $k$-NN instance's complementary label set. Let $q$ denote the probability of the label $y'$ appearing in its $k$-NN instance's complementary label set. The NLE error rate $\epsilon_2$ is upper bounded by:*

$$\epsilon_2 = \mathbb{P}(y_i \in \widehat{\bar{Y}}_i) \leq \sum\nolimits_{j=1}^{k} \binom{|Y_i|-1}{|Y_i|-\tau_i} F_{\beta_k}(k-j+1,j)^{(|Y_i|-\tau_i)} b_{\alpha_k}(k,j), \tag{22}$$

where $F_{\beta_k}(k,j) = \int_0^{\beta_k} p^{k-1}(1-p)^{j-1}dt$ denotes the regularized incomplete beta function, $b_{\alpha_k}(k,j) = \binom{k}{j}\alpha_k^j(1-\alpha_k)^{k-j}$ is the probability mass function of a binomial distribution $B(k,\alpha_k)$. Detailed proofs are provided in Appendix D.

***Remark.*** Theorem 2 and Theorem 3 show that both PLG error rate $\epsilon_1$ and NLE error rate $\epsilon_2$ are upper bounded under mild condition.

Table 1: Comparison of classification accuracies between different methods on four datasets with a single complementary label per instance. The results (mean $\pm$ std) are reported over 3 random trials. The best results are highlighted in bold (The same applies hereinafter).

| Method | STL-10 | SVHN | FMNIST | CIFAR-10 |
|---|---|---|---|---|
| UB-EXP | 28.84±0.54% | 88.93±0.17% | 87.96±0.08% | 62.90±0.06% |
| UB-LOG | 20.41±0.46% | 89.59±0.08% | 87.59±0.14% | 70.28±0.12% |
| SCL-EXP | 31.03±0.61% | 88.66±0.20% | 88.31±0.09% | 72.35±0.10% |
| SCL-LOG | 30.74±0.72% | 89.26±0.24% | 88.03±0.10% | 79.87±0.14% |
| POCR | 34.96±0.32% | 96.65±0.14% | 92.29±0.07% | 94.15±0.09% |
| SELF-CL | 30.87±0.72% | 90.13±0.23% | 84.86±0.10% | 88.95±0.22% |
| ComCo | 32.43±0.28% | 91.41±0.35% | 85.42±0.40% | 89.36±0.76% |
| Ours | **55.25±0.36%** | **97.58±0.18%** | **93.38±0.06%** | **94.78±0.12%** |

Table 2: Comparison of classification accuracies between different methods on five datasets with multiple complementary labels per instance. The results (mean $\pm$ std) are reported over 3 random trials.

| Method | STL-10 | SVHN | FMNIST | CIFAR-10 | CIFAR-100 |
|---|---|---|---|---|---|
| UB-EXP | 60.85±0.12% | 95.23±0.09% | 92.34±0.28% | 91.13±0.23% | 34.43±0.08% |
| UB-LOG | 62.84±0.17% | 94.76±0.07% | 91.84±0.29% | 92.01±0.21% | 52.76±0.15% |
| SCL-EXP | 62.96±0.10% | 95.28±0.14% | 92.20±0.27% | 91.85±0.25% | 47.81±0.09% |
| SCL-LOG | 61.60±0.14% | 94.88±0.16% | 91.51±0.25% | 92.67±0.18% | 49.40±0.19% |
| POCR | 74.51±0.29% | 97.14±0.09% | 94.76±0.26% | 96.09±0.27% | 53.16±0.11% |
| SELF-CL | 69.85±0.20% | 91.58±0.30% | 94.92±0.21% | 92.23±0.16% | 57.65±0.25% |
| ComCo | 73.28±0.19% | 95.41±0.23% | 92.01±0.16% | 91.38±0.73% | 57.88±0.95% |
| Ours | **77.11±0.14%** | **98.13±0.11%** | **95.16±0.13%** | **96.80±0.28%** | **64.33±0.43%** |

# 6 EXPERIMENT

## 6.1 EXPERIMENT SETUP

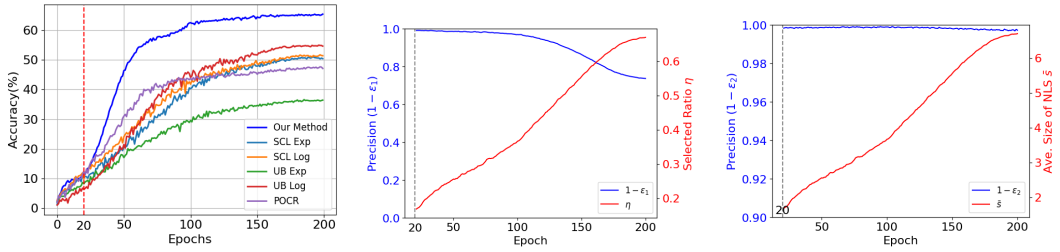
**Datasets.** We use five commonly used benchmark datasets, STL-10 (Coates et al., 2011), Fashion-MNIST (FMNIST) (Xiao et al., 2017), SVHN (Netzer et al., 2011), CIFAR-10 and CIFAR-100 (Krizhevsky and Hinton, 2009). We conduct experiments by considering both the scenarios of Single CLL (SCLL) and Multiple CLL (MCLL). To generate single complementary label, we randomly select one of the complementary classes per instance. To generate multiple complementary labels, let $s$ be the size of $\bar{Y}$, we first instantiate $p(s) = \binom{K-1}{s}/(2^K-2), s \in \{1,2,...,K-1\}$, which represents the possibility of randomly sample a complementary label set whose size is $s$ from all possible complementary label sets which has $2^K-2$ sets to choose from. Then for each instance

$\boldsymbol{x}_i$, we first sample $s_i$ from $p(s_i)$, and then sample a complementary label set $\bar{Y}_i$ with size $s$ from $p(\bar{Y}_i) = 1/\binom{K-1}{s_i}$.

**Compared methods.** We compare the performance of PLNL with seven state-of-the-art CLL methods, including UB-EXP (Feng et al., 2020), UB-LOG (Feng et al., 2020), SCL-EXP (Chou et al., 2020), SCL-LOG (Chou et al., 2020), POCR (Wang et al., 2021), SELF-CL (Liu et al., 2022) and ComCo (Jiang et al., 2024) and two state-of-the-art SSL methods, Fixmatch (Sohn et al., 2020) and Freematch (Wang et al., 2022).

**Implementation.** Implementation details are provided in Appendix F.

### 6.2 Main Results




(a) Accuracy Comparison of CLL methods (b) PLG Precison vs. Selected Ratio (c) NLE Precison vs. Ave. Num. of NLS

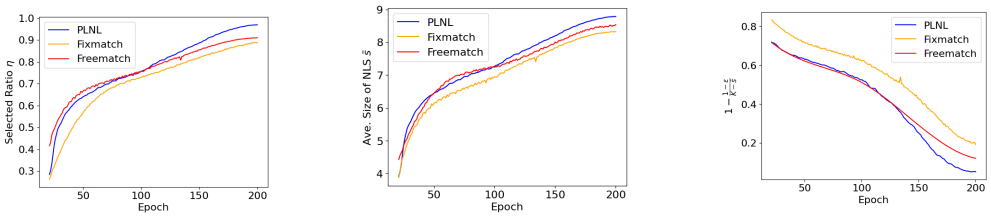
Figure 2: The experiments is conducted on CIFAR-100 with multiple complementary labels (MCLL). (a) The accuracy of PLNL improves tremendously over epochs and achieves the best finally. (b) The precision of PLG decreases slowly, while the selected ratio steadily rises, indicating a growing proportion of selected instances during training. (c) Ave. Size of NLS denotes average number of negative label set $\bar{s}$. The precision of NLE remains relatively stable with a slight decrease, whereas the average size of negative label set increases significantly, showing a steady recovery of negative labels.

**PLNL achieves SOTA results.** As shown in Table 1 and Table 9, PLNL outperforms all the compared method by a significant margin across all datasets. Specifically, on STL-10 dataset, we outperform the previous SOTA by **20.29%** and **1.61%** in both SCLL and MCLL settings. Furthermore, PLNL performs even better in harder scenarios where there is larger label space or less supervised information for each class. We challenge this by showing our results on CIFAR-100 datasets. On CIFAR-100 with Multiple CLs, the improvement is **6.45%** compared to previous SOTA. Fig. 2a further demonstrates that PLNL significantly outperforms the compared ones.

**PLNL pseudo-labeling achieves excellent performance with extremely high precision.** Fig. 2b shows that as the number of epochs increases, PLG will select more and more highly-confident instance, eventually occupying most of the dataset, while the precision only drops slightly in the final stage, which maintains high precision and high selected ratio. Meanwhile, Fig. 2c shows that NLE identifies more and more negative labels with extremely high precision, which maintains above 0.99 throughout training.

**Compared with SOTA semi-supervised learning methods, PLNL performs even better.** From Fig. 3a and Fig. 3b, we can see that PLNL achieves both higher selected ratio and recovered more negative labels compared with Fixmatch and Freematch. PLNL is both accurate and comprehensive in recovering label information. This highlights PLNL's enhanced capacity in leveraging moderately-confident and under-confident instances for label recovery, showcasing both stability and scalability.

**PLNL can reduce the generalization bound and achieves lower generalization bound compared with SSL methods.** Fig. 3c demonstrates that $(1 - \frac{1-\epsilon}{K-\bar{s}})$ decreases as the training progresses. As $(1 - \frac{1-\epsilon}{K-\bar{s}})$ is the variable of the generalization bound derived in Theorem 1, it is safe to conclude that PLNL can continuously and significantly reduce the generalization error.




(a) Selected Ratio vs. Epoch (b) Ave. Num. of NLS vs. Epoch (c) Convergence of $(1-\frac{1-\epsilon}{K-\bar{s}})$

Figure 3: The experiments is conducted on CIFAR-10 with single complementary labels (SCLL). (a) shows that selected ratio of PLNL trenscends Fixmatch and Freematch significantly. (b) shows that average size of NLS of PLNL is significantly larger due to specially designed technique NLE for enhancing the untrustworthy negative labels. Nearly all negative labels are revealed at the end of training, almost reaching 9 negative labels for each instances in CIFAR-10 (c) indicates that the value of $(1-\frac{1-\epsilon}{K-\bar{s}})$ decreases steadily during training.

### 6.3 Ablation Study

**Two-view networks facilitate increased pseudo-labeling precision.** We observe that two-view network significantly boost the performance of PLNL pseudo-labeling, which helps accurately select more highly-confident instances. As shown in Table 3, the $\eta$ increases **8.75%** and **8.65%** respectively on CIFAR-10 and CIFAR-100 with $1-\epsilon_1$ increases **6.58%** and **9.22%** respectively. We also compare PLNL with one variant: PLNL v1 where we replace the two-view networks in PLNL with a single network in Table 4, which shows that ours outperforms PLNL v1 by a remarkable margin (e.g. **+6.00%** on STL-10).

Table 3: The performance of PLNL with single network on two settings.

| Method | CIFAR-10 SCLL | | CIFAR-100 MCLL | |
|---|---|---|---|---|
| | $\eta$ | 1-$\epsilon_1$ | $\eta$ | 1-$\epsilon_1$ |
| Single | 84.65 | 90.21 | 67.43 | 70.62 |
| Two-view | **93.40** | **96.79** | **76.08** | **79.84** |

**All the components contribute to the performance gain.** We explore the effectiveness of our proposed PLG and NLE method on three settings STL-10 (SCLL), CIFAR-10 (SCLL), CIFAR-100 (MCLL). Specifically, we compare PLNL with several variants: (1) PLNL v2 which removes the PLG component; (2) PLNL v3 which removes the NLE component. From Table 4, we observe that PLNL outperforms PLNL v2 remarkably (e.g., **+5.43%** on STL-10), which proves the effectiveness of positive label guessing. We also observe that ours outperforms PLNL v3 (e.g., **+1.19%** on CIFAR-100), which proves the effectiveness of NLE.

Table 4: Classification accuracy of degenerated methods on three settings.

| Method | STL-10 SCLL | CIFAR-10 SCLL | CIFAR-100 MCLL |
|---|---|---|---|
| PLNL | **55.25** | **94.78** | **64.33** |
| PLNL v1 | 49.25 | 93.75 | 63.09 |
| PLNL v2 | 49.82 | 92.01 | 58.94 |
| PLNL v3 | 53.22 | 94.28 | 63.14 |
| POCR | 34.96 | 94.15 | 53.16 |

## 7 Conclusion.

In this paper, we introduce a novel complementary label learning method that reformulates CLL as an inverse problem to infer the full label information from the output space information. To this end, we split this inverse problem into two subtasks (PLG and NLE). A confidence-based instances selection module is proposed for dataset split: highly-confident, moderately-confident and under-confident. Then we perform PLG for highly-confident instances by assigning pseudo-labels to them. For moderately-confident and under-confident instances, we perform NLE by enhancing their negative label set with different levels and train them with the augmented negative labels iteratively. We theoretically prove that when pseudo-labeling error is limited, we can construct a classifier consistent with that learned by clean full labels. The upper bounds of PLG and NLE error rate are deduced and we empirically show that PLNL can infer both positive and negative labels with a high precision. We conducted extensive experiments which demonstrate that PLNL achieves a new state-of-the-art in CLL. In addition, extensive ablation studies have proved the effectiveness of each component.

## REFERENCES


Shoshana Abramovich, Graham Jameson, and Gord Sinnamon. Refining jensen's inequality. *Bulletin mathématique de la Société des Sciences Mathématiques de Roumanie*, pages 3–14, 2004.

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *International conference on machine learning*, pages 1929–1938. PMLR, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

Lei Feng and Bo An. Partial label learning with self-guided retraining. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3542–3549, 2019.

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *International conference on machine learning*, pages 3072–3081. PMLR, 2020.

Xiuwen Gong, Dong Yuan, Wei Bao, and Fulin Luo. A unifying probabilistic framework for partially labeled data learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Shuo He, Chaojie Wang, Guowu Yang, and Lei Feng. Candidate label set pruning: A data-centric perspective for deep partial-label learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Fk5IzauJ7F.

Jintao Huang and Yiu-Ming Cheung. Trustworthy partial label learning with out-of-distribution detection. *arXiv preprint arXiv:2403.06681*, 2024.

Zhizhong Huang, Junping Zhang, and Hongming Shan. Twin contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11661–11670, 2023.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. *Advances in neural information processing systems*, 30, 2017.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *International conference on machine learning*, pages 2971–2980. PMLR, 2019.

Hiroki Ishiguro, Takashi Ishida, and Masashi Sugiyama. Learning from noisy complementary labels with robust loss functions. *IEICE TRANSACTIONS on Information and Systems*, 105(2):364–376, 2022.

Haoran Jiang, Zhihao Sun, and Yingjie Tian. Comco: Complementary supervised contrastive learning for complementary label learning. *Neural Networks*, 169:44–56, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Takuo Kaneko, Issei Sato, and Masashi Sugiyama. Online multiclass classification based on prediction margin for partial feedback. *arXiv preprint arXiv:1902.01056*, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023a. URL https://arxiv.org/abs/2301.12597.

Siyuan Li, Weiyang Jin, Zedong Wang, Fang Wu, Zicheng Liu, Cheng Tan, and Stan Z Li. Semireward: A general reward model for semi-supervised learning. *arXiv preprint arXiv:2310.03013*, 2023b.

Jiabin Liu, Biao Li, Minglong Lei, and Yong Shi. Self-supervised knowledge distillation for complementary label learning. *Neural Networks*, 155:318–327, 2022.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *International conference on machine learning*, pages 6500–6510. PMLR, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, Spain, 2011.

Gang Niu, Marthinus Christoffel Du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. *Advances in neural information processing systems*, 29, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

Mina Rezaei, Haojin Yang, and Christoph Meinel. Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation. *Multimedia Tools and Applications*, 79(21):15329–15348, 2020.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 2022.

William Tapper, Gustavo Carneiro, Christos Mikropoulos, Spencer A Thomas, Philip M Evans, and Stergios Boussios. The application of radiomics and ai to molecular imaging for prostate cancer. *Journal of personalized medicine*, 14(3):287, 2024.

Shiyu Tian, Hongxin Wei, Yiqun Wang, and Lei Feng. Crosel: Cross selection of confident pseudo labels for partial-label learning, 2024.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Learning from complementary labels via partial-output consistency regularization. In *IJCAI*, pages 3075–3081, 2021.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13726–13735, 2020.

Shiyu Xia, Jiaqi Lv, Ning Xu, Gang Niu, and Xin Geng. Towards effective visual representations for partial-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15589–15598, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Kouzhiqiang Yucheng Xie, Jing Wang, Yuheng Jia, Boyu Shi, and Xin Geng. Rankmatch: A novel approach to semi-supervised label distribution learning leveraging inter-label correlations. *arXiv preprint arXiv:2312.06343*, 2023.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, 2022.

Manyi Zhang, Xuyang Zhao, Jun Yao, Chun Yuan, and Weiran Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15890–15900, 2023.

# A THE PSEUDO-CODE OF PLNL

**Algorithm 1** Pseudo-code of PLNL.

**Input:** Training dataset $\mathcal{D}$, mini-batch size $B$, epochs $E_{max}$, a two-view network with shared parameter $\Theta$ and two memory bank matrices $\boldsymbol{M}^w$, $\boldsymbol{M}^s$, hyperparameters $t$, $k$, $\tau_m$, $\tau_u$.

1: Initialize $\boldsymbol{M}^w$, $\boldsymbol{M}^s$, $\Theta$ by warming up 20 epochs using SCL-LOG (Chou et al., 2020).
2: **for** $e = 1, 2, \ldots, E_{max}$ **do**
3: Shuffle $\mathcal{D}$ into $\frac{|\mathcal{D}|}{B}$ mini-batches;
4: Construct $\boldsymbol{M}^w$, $\boldsymbol{M}^s$;
5: Compute two-view network outputs of each instances;
6: Select $\mathcal{H}, \mathcal{M}, \mathcal{U}$ based on criteria in Eq. (5);
7: Pseudo-label the highly-confident set $\mathcal{H}$; // PLG
8: Enhance negative labels of set $\mathcal{M}, \mathcal{U}$ based on Eq. (14) and Eq. (15); // NLE
9: **for** $i = 1, 2, \ldots, \frac{|\mathcal{D}|}{B}$ **do**
10: Fetch an batch $B_i$ with enhanced negative label set;
11: Compute the empirical risk $\widehat{R}'(f)$ by Eq. (19);
12: Update parameters of $\Theta$;
13: **end for**
14: **end for**

**Output:** parameters of $\Theta$.

# B PROOF OF THEOREM 1

We first derive the uniform deviation bound between $\widehat{R}(f)$ and $R(f)$.

**Lemma 1.** *Suppose that the binary loss function $\ell(f(\boldsymbol{x}), y)$ is $\rho$-Lipschitz continuous* w.r.t. $f(\boldsymbol{x})$. *For any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\left| R(f) - \widehat{R}(f) \right| \leq 2\rho K \mathfrak{R}_N(\mathcal{F}) + KB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}. \tag{23}$$

*Proof.* We firstly define the Rademacher complexity of $\mathcal{L}$ and $\mathcal{F}$ with $N$ training instances as follows:

$$\begin{aligned} &\mathfrak{R}_N(\mathcal{L} \circ \mathcal{F}) \\ &= \mathbb{E}_{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^{N} \sigma_i \mathcal{L}_{CLL}(f(\boldsymbol{x}_i), \widehat{\bar{Y}}_i), ) \right]. \end{aligned} \tag{24}$$

Considering that $\mathcal{L}_{CLL}(f(\boldsymbol{x}), \widehat{\bar{Y}}_i) = \sum_{y \notin \widehat{\bar{Y}}_i} \ell(f(\boldsymbol{x}), y)$, we have

$$\begin{aligned} \mathfrak{R}_N(\mathcal{L} \circ \mathcal{F}) &\leq K \mathfrak{R}_N(\ell \circ \mathcal{F}), \\ &\leq \rho K \mathfrak{R}_N(\mathcal{F}), \end{aligned} \tag{25}$$

where the second line is based on Lipschitz continuity of $\ell(f(\boldsymbol{x}), y)$.

Suppose an instance $(\boldsymbol{x}_i, y_i)$ is replaced by another arbitrary instance $(\boldsymbol{x}_i^{'}, y_i^{'})$, this leads to a change of $\sup_{f \in \mathcal{F}} R(f) - \widehat{R}(f)$ no greater than $\frac{KB}{N}$ due to the fact that $\ell$ is bounded by $B$. According to McDiarmid's inequality (Mohri et al., 2018), for any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$, we have

$$\sup_{f \in \mathcal{F}} R(f) - \widehat{R}(f) \leq \mathbb{E}\left[ \sup_{f \in \mathcal{F}} R(f) - \widehat{R}(f) \right] + KB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}. \tag{26}$$

In addition, it is routine (Mohri et al., 2018) that

$$\mathbb{E}\left[ \sup_{f \in \mathcal{F}} R(f) - \widehat{R}(f) \right] \leq 2\bar{\mathfrak{R}}_N(\mathcal{F}). \tag{27}$$

By combining Eq. (26) and Eq. (27), and futher taking the other direction of $\sup_{f\in\mathcal{F}} \widehat{R}(f) - R(f)$ into account, with probability at least $1-\delta$, we have

$$\sup_{f\in\mathcal{F}} \left|R(f) - \widehat{R}(f)\right| \leq 2\rho K \mathfrak{R}_N(\mathcal{F}) + KB\sqrt{\frac{\log\frac{2}{\delta}}{2N}}, \tag{28}$$

which concludes the proof.

Due to inevitable error made in PLG and NLE. The problem of CLL actually has been translated into a noisy CLL problem where the true label might be mislabeled as complementary label (Ishiguro et al., 2022). Let $\epsilon_1$ be the error rate of PLG and $\epsilon_2$ be the error rate of NLE. Since NLE error rate $\epsilon_1 = \sum_{i=1}^{N_h} \frac{\mathbb{I}(y_i \in \widehat{\bar{Y}}_i)}{N_h}$ and NLE error rate $\epsilon_2 = \sum_{i=1}^{N_m+N_u} \frac{\mathbb{I}(y_i \in \widehat{\bar{Y}}_i)}{N_m+N_u}$, the actual noise rate $\epsilon = \frac{\sum_{i=1}^{N} \mathbb{I}(y_i \in \widehat{\bar{Y}}_i)}{N}$ can be calculated as $\epsilon = \frac{N_h}{N}\epsilon_1 + \frac{N_m+N_u}{N}\epsilon_2 = \eta\epsilon_1 + (1-\eta)\epsilon_2$. We further bound the difference between $\widehat{R}(f)$ and $\hat{R}'(f)$.

**Lemma 2.** *Suppose that the binary loss function $\ell$ is bounded by $B$. For some noise rate $\epsilon \in (0,1)$ and average complementary label size $\bar{s}$ for any $f \in \mathcal{F}$, we have*

$$\left|\widehat{R}'(f) - \widehat{R}(f)\right| \leq (1 - \frac{1-\epsilon}{K-\bar{s}})B. \tag{29}$$

We firstly proved the upper bound of the $\widehat{R}'(f)$:

$$\begin{aligned}
\widehat{R}'(f) &= \frac{1}{N}\sum_{i=1}^{N} \bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i), \widehat{\bar{Y}}_i), \\
&= \frac{1}{N}\sum_{i=1}^{N} \ell(f(\boldsymbol{x}_i), y_i) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \notin \widehat{\bar{Y}}_i) \left[\sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i} \frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i), c) - \frac{K-|\widehat{\bar{Y}}_i|-1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i), y_i)\right] \\
&+ \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \in \widehat{\bar{Y}}_i) \left[\sum_{c\notin\widehat{\bar{Y}}_i} \frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i), c) - \ell(f(\boldsymbol{x}_i), y_i)\right], \\
&\leq \widehat{R}(f) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \notin \widehat{\bar{Y}}_i) \sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i} \frac{1}{K-\bar{s}}\ell(f(\boldsymbol{x}_i), c) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \in \widehat{\bar{Y}}_i) \sum_{c\notin\widehat{\bar{Y}}_i} \frac{1}{K-\bar{s}}\ell(f(\boldsymbol{x}_i), c), \\
&\leq \widehat{R}(f) + (1-\epsilon)\frac{K-\bar{s}-1}{K-\bar{s}}B + \epsilon B,
\end{aligned} \tag{30}$$

where the second line holds based on $\epsilon = \frac{\sum_{i=1}^{N} \mathbb{I}(y_i \in \widehat{\bar{Y}}_i)}{N}$ and Jensen's inequality (Abramovich et al., 2004) as , and we can prove the lower bound in a similar way:

$$\begin{aligned}
\widehat{R}'(f) &= \frac{1}{N}\sum_{i=1}^{N} \ell(f(\boldsymbol{x}_i), y_i) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \notin \widehat{\bar{Y}}_i) \left[\sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i} \frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i), c) - \frac{K-|\widehat{\bar{Y}}_i|-1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i), y_i)\right] \\
&+ \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \in \widehat{\bar{Y}}_i) \left[\sum_{c\notin\widehat{\bar{Y}}_i} \frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i), c) - \ell(f(\boldsymbol{x}_i), y_i)\right], \\
&\geq \widehat{R}(f) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \notin \widehat{\bar{Y}}_i) \sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i} \frac{1}{K-\bar{s}}\ell(f(\boldsymbol{x}_i), c) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i \in \widehat{\bar{Y}}_i) \sum_{c\notin\widehat{\bar{Y}}_i} \frac{1}{K-\bar{s}}\ell(f(\boldsymbol{x}_i), c), \\
&\geq \widehat{R}(f) - (1-\epsilon)\frac{K-\bar{s}-1}{K-\bar{s}}B - \epsilon B,
\end{aligned} \tag{31}$$

By combining these two sides, we have:

$$\left|\widehat{R}'_h(f) - \widehat{R}_h(f)\right| \leq 1 - \frac{1-\epsilon}{K-\bar{s}}, \tag{32}$$

which concludes the proof.

Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
& R(\hat{f}) \\
& \leq \widehat{R}(\hat{f}) + 2\rho K \mathfrak{R}_N(\mathcal{F}) + KB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \\
& \leq \widehat{R}'(\hat{f}) + (1 - \frac{1-\epsilon}{K-\bar{s}})B + 2\rho K \mathfrak{R}_N(\mathcal{F}) + KB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \\
& \leq \widehat{R}'(f^*) + (1 - \frac{1-\epsilon}{K-\bar{s}})B + 2\rho K \mathfrak{R}_N(\mathcal{F}) + KB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \\
& \leq \widehat{R}(f^*) + 2(1 - \frac{1-\epsilon}{K-\bar{s}})B + 2\rho K \mathfrak{R}_N(\mathcal{F}) + KB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \\
& \leq R(f^*) + 2(1 - \frac{1-\epsilon}{K-\bar{s}})B + 4\rho K \mathfrak{R}_N(\mathcal{F}) + 2KB\sqrt{\frac{\log \frac{2}{\delta}}{2N}},
\end{aligned} \tag{33}
$$

where the first and fifth line are based on Lemma 1, the second and fourth line are based on Lemma 2, the third line is based on the definition of the empirical risk minimizer $\hat{f} = \text{argmin}_{f \in \mathcal{F}} \widehat{R}(f)$ which means any other $f \neq \hat{f}$ would lead to a larger risk of $\widehat{R}(f)$.

## C Proof of Theorem 2

Motivated by the formulation of partially labeled data learning in (Gong et al., 2022), we assume that the full label information $Y = \{y, \widetilde{Y}\}$ where $y$ is the ground-truth label and $\widetilde{Y}$ is a set of the rest labels. And we assume that $Q$ is a hypothetical predictive model with parameters $\theta$, where $\theta$ is the model parameter used for prediction. PLG aims to identify the ground-truth label from the label set except complementary label set, which can be implemented by maximizing the conditional likelihood of training dataset with respect to parameters $\theta$. The conditional log-likelihood given all training examples can be expressed as follows.

$$
f(\theta|\widetilde{Y}) = \frac{1}{N}\sum_{i=1}^{N} \log Q(y_i|\boldsymbol{x}_i). \tag{34}
$$

By multiplying and dividing classifier Q by the true distribution of identified ground-truth labels given features $P(y|\boldsymbol{x})$, we can formulate Eq. (34) as follows.

$$
f(\theta|\widetilde{Y}) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{Q(y_i|\boldsymbol{x}_i)}{P(y_i|\boldsymbol{x}_i)} + \frac{1}{N}\sum_{i=1}^{N} \log P(y_i|\boldsymbol{x_i}). \tag{35}
$$

By multiplying and dividing the probability $P(Y|\boldsymbol{x})$ to the second term, we can formulate Eq. (34) as follows.

$$
f(\theta|\widetilde{Y}) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{Q(y_i|\boldsymbol{x}_i)}{P(y_i|\boldsymbol{x_i})} + \frac{1}{N}\sum_{i=1}^{N} \log \frac{P(y_i|\boldsymbol{x_i})}{P(Y_i|\boldsymbol{x_i})} + \frac{1}{N}\sum_{i=1}^{N} \log P(Y_i|\boldsymbol{x_i}). \tag{36}
$$

We use $\mathbb{E}_{(\mathcal{X},\mathcal{Y})}$ operator to calculate the expectation of the random variables $(\mathcal{X}, \mathcal{Y})$, meaning $n \to \infty$.

$$
f(\theta|\widetilde{Y}) = \mathbb{E}_{(\mathcal{X},\mathcal{Y})}\left\{\log \frac{Q(y|\boldsymbol{x})}{P(y|\boldsymbol{x})}\right\} + \mathbb{E}_{(\mathcal{X},\mathcal{Y})}\left\{\log \frac{P(y|\boldsymbol{x})}{P(Y|\boldsymbol{x})}\right\} + \mathbb{E}_{(\mathcal{X},\mathcal{Y})}\left\{\log P(Y|\boldsymbol{x})\right\}. \tag{37}
$$

Recall that we assume that the full label information $Y = \{y, \widetilde{Y}\}$. Then the second term of Eq. (37) can be developed as follows.

$$
\begin{aligned}
\mathbb{E}_{(\mathcal{X},\mathcal{Y})}\left\{\log \frac{P(y|\boldsymbol{x})}{P(Y|\boldsymbol{x})}\right\} &= -\mathbb{E}_{(\mathcal{X},\mathcal{Y})}\left\{\log \frac{P(Y|\boldsymbol{x})}{P(y|\boldsymbol{x})}\right\}\},\\
&= -\sum_{(\boldsymbol{x},Y)} P(\boldsymbol{x},Y)\log\frac{P(y,\widetilde{Y}|\boldsymbol{x})}{P(y|\boldsymbol{x})},\\
&= -\sum_{(\boldsymbol{x},Y)} P(\boldsymbol{x},Y)\log\frac{P(y,\widetilde{Y},\boldsymbol{x})}{P(y,\boldsymbol{x})},\\
&= -\sum_{(\boldsymbol{x},Y)} P(\boldsymbol{x},Y)\log\frac{P(\widetilde{Y},\boldsymbol{x}|y)}{P(\boldsymbol{x}|y)},\\
&= -\sum_{(\boldsymbol{x},Y)} P(\boldsymbol{x},Y)\log\frac{P(\widetilde{Y},\boldsymbol{x}|y)}{P(\boldsymbol{x}|y)}\frac{P(\widetilde{Y}|y)}{P(\widetilde{Y}|y)},\\
&= -\sum_{(\boldsymbol{x},Y)} P(\boldsymbol{x},Y)\log\frac{P(\widetilde{Y},\boldsymbol{x}|y)}{P(\boldsymbol{x}|y)P(\widetilde{Y}|y)} - \sum_{(\boldsymbol{x},Y)} P(\boldsymbol{x},Y)\log P(\widetilde{Y}|y),\\
&= -\sum_{(\boldsymbol{x},Y)} P(\boldsymbol{x},Y)\log\frac{P(\widetilde{Y},\boldsymbol{x}|y)}{P(\boldsymbol{x}|y)P(\widetilde{Y}|y)} - \sum_{(\widetilde{Y},y)} P(\widetilde{Y},y)\log P(\widetilde{Y}|y),\\
&= -I(\widetilde{Y},X|y) + H(\widetilde{Y}|y),\\
&= -I(\widetilde{Y},X|y). \qquad (38)
\end{aligned}
$$

The last equality holds because the conditional entropy $H(\widetilde{Y}|y) = 0$. This is because in CLL setting, once $y$ is known, then the full label information is of course known in advance, meanwhile, thus the uncertainty remaining in $\widetilde{Y}$ is zero, i.e., $H(\widetilde{Y}|y) = 0$. By combining Eq. (38) and Eq. (37), we have the objective function as follows.

$$
f(\theta|\widetilde{Y}) = -\mathbb{E}_{(\mathcal{X},\mathcal{Y})}\left\{\log\frac{P(y|\boldsymbol{x})}{Q(y|\boldsymbol{x})}\right\} - I(\widetilde{Y},X|y) - H(Y|X). \qquad (39)
$$

The first term is a log likelihood ratio between the true and the predicted ground-truth label distributions given features. The value of this term depends on how well the model $Q$ can approximate $P$. The second term is the conditional mutual information between the complementary labels and the features, given the ground-truth label. The last term is a constant independent of parameters.

Then, we discuss the mild assumption under which PLG method is effective for CLL.

**Assumption 2.** *Let $y \in Y'$ and $y' \in Y'$ denote any ground-truth label and unidentified negative label respectively. Let $X$ denote the random variables of $\boldsymbol{x}$. Let $I(y, X)$ denote the mutual information between ground-truth label $y$ and feature $X$. Let $I(y', X)$ denote the mutual information between any unidentified negative label $y'$ and feature $X$. Then, with probability no more than $\psi$, we have*

$$
I(y, X) \leq I(y', X), \qquad (40)
$$

where $\psi < \frac{1}{K-1-s}$.

*Remark.* This assumption ensures that the feature tends to have more mutual information with positive labels than negative labels.

Motivated by the simplification for identification method in (Gong et al., 2022). In PLG, the training objective actually is the second term of Eq. (39):$f_y(\boldsymbol{x}) = I(y, X|\widetilde{Y})$. As a result, the error rate of PLG can be calculated as follows. Suppose $\hat{y}$ is the guessed positive label, and $y'$ is any negative

label.

$$\begin{aligned}\mathbb{P}(y \neq j, \hat{y} = j) &= \sum\nolimits_{y' \notin \bar{Y}} P(f_{\hat{y}}(\boldsymbol{x}) - f_{y'}(\boldsymbol{x}) \leq 0), \\ &= \sum\nolimits_{y' \notin \bar{Y}} P(I(\hat{y}, X|\widetilde{Y}) \leq I(y', X|\widetilde{Y})), \\ &\leq (K - 1 - s)\psi,\end{aligned} \tag{41}$$

where the second line is based on replacing $\widetilde{Y}$ with any $y$ in 39, the fourth line is based on assumption 2. This concludes the proof.

## D Proof of Theorem 3

It is evident that reliable representation information is of great importance to the performace of $k$-NN based NLE. Motivated by the label distinguishability setting in (He et al., 2024), a mild assumption for CLL datasets are discussed to ensure the reliability of the representation information.

According to assumption 1, we have the following lemma.

**Lemma 3.** *$\forall(\boldsymbol{x}_i, \bar{Y}_i) \in \mathcal{D}$, let $p$ denote the probability of the true label $y_i \in Y_i$ appearing in its $k$-NN instance's complementary label set. Let $q$ denote the probability of each non-complementary negative label $y \in Y_i \backslash \{y_i\}$ appearing in its $k$-NN instance's complementary label set. Then, we have*

$$p \leq \alpha_k, q \geq \beta_k. \tag{42}$$

Then, We derive the error bound of NLE in a step by step manner as follows.

$$\begin{aligned}&\mathbb{P}(\boldsymbol{F}_i^{(\tau)} < \boldsymbol{F}_{iy}) = \sum_{j=0}^{k} P(\boldsymbol{F}_i^{(\tau)} < \boldsymbol{F}_{iy} | \boldsymbol{F}_{iy} = j) P(\boldsymbol{F}_{iy} = j) \\ &= \sum_{j=1}^{k} P(\boldsymbol{F}_i \leq j-1) P(\boldsymbol{F}_i = j), \\ &= \sum_{j=1}^{k} \left( \underbrace{P(\boldsymbol{F}_i^{(1)}) > j) \ldots}_{(\tau-1)\textit{items}} \underbrace{P(\boldsymbol{F}_i^{(\tau)}) < j) \ldots P(\boldsymbol{F}_i^{(|Y_i|-1)}) < j)}_{(|Y_i|-\tau)\textit{items}} \right) P(\boldsymbol{F}_i) = j), \\ &\leq \sum\nolimits_{j=1}^{k} \binom{|Y_i|-1}{|Y_i|-\tau} F_{\beta_k}(k-j+1, j)^{(|Y_i|-\tau)} b_{\alpha_k}(k, j),\end{aligned} \tag{43}$$

where $F_{\beta_k}(k, j) = \int_0^{\beta_k} p^{k-1}(1-p)^{j-1} dt$ denotes the regularized incomplete beta function. $b_{\alpha_k}(k, j) = \binom{k}{j} \alpha_k^j (1 - \alpha_k)^{k-j}$ is simply the probability mass function of a binomial distribution $B(k, \alpha_k)$ which is used to describe the counting process of the label frequency $\boldsymbol{F}_{iy}$.

## E Details of MCLL settings.

Ishida *et al.* (Ishida et al., 2017) assumed that $\bar{p}(\boldsymbol{x}, \bar{y})$ is expressed as:

$$\bar{p}(\boldsymbol{x}, \bar{y}) = \frac{1}{c-1} \sum\nolimits_{y \neq \bar{y}} p(\boldsymbol{x}, y). \tag{44}$$

This assumption implies that all other labels except the correct label are picked as the complementary label with uniform probabilities. Later, Feng *et al.* (Feng et al., 2020) considered a more general setting where each instance is associated with multiple complementary labels (Multiple CLL). Suppose a Multiple CLL dataset is represented as $\{(\boldsymbol{x}_i, \bar{Y}_i)\}_{i=1}^N$, where $\bar{Y}_i$ is the complementary label set for instance $\boldsymbol{x}_i$. Let us denote the number of the complementary labels by a random variable $s_i$, which is sampled from a distribution $p(s_i)$. Then, we assume that each sample is drawn from the following distribution:

$$\bar{p}(\boldsymbol{x}_i, \bar{Y}_i) = \sum\nolimits_{j=1}^{c-1} \bar{p}(\boldsymbol{x}_i, \bar{Y}_i | s_i = j) p(s_i = j), \tag{45}$$

where

$$\bar{p}(\boldsymbol{x}_i, \bar{Y}_i | s_i = j) p(s_i = j) := \left\{ \begin{array}{ll} \frac{1}{\binom{c-1}{j}} \sum_{y \notin \bar{Y}_i} p(\boldsymbol{x}, y), & \text{if} s_i = j, \\ 0, & \text{otherwise.} \end{array} \right\} \tag{46}$$

# F MORE EXPERIMENT DETAILS

**Details of Compared Methods.**

- UB-EXP (Feng et al., 2020), an unbiased risk estimator with an estimation error bound, which is derived for Multiple CLL specially.
- UB-LOG (Feng et al., 2020), another unbiased risk estimator with an estimation error bound but with a different multi-class classification loss function.
- SCL-EXP (Chou et al., 2020), a surrogate complementary loss with the use of exponential loss function.
- SCL-LOG (Chou et al., 2020), a surrogate complementary loss with the use of negative log loss function.
- POCR (Wang et al., 2021), an algorithm which combines the SCL-LOG loss and the consistency regularization technique.
- SELF-CL (Liu et al., 2022), an self-supervised learning algorithm which integrates self-distillation to CLL.
- ComCo (Jiang et al., 2024), a contrastive learning framework which leverages the contrastive learning technique on CLL.

**Details of Implementation.**

**Implementation.** Values of hyperparameters in PLNL are set as follows. The queue size $t$ is selected from $\{2, 3, 4, 5\}$, $k$-NN parameter $k$ is selected from $\{100, 250, 500\}$. The $\alpha$ in the instance-aware self-adaptive threshold is selected from $\{0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. For each method, we train the commonly used PreAct-ResNet18 (He et al., 2016) with 200 epochs (initial 20 epochs for warm-up), and use SGD as the opimizer with a momentum of 0.9, a weight decay of 1e-4. We set the batch size from $\{64, 128\}$, the initial learning rate from $\{10^{-1}, 10^{-2}\}$, and we use cosine learning rate scheduling with final learning rate $10^{-3}$. We employed *faiss* (Johnson et al., 2019) to compute $k$-NN instances in the output space, which is a library for efficient similarity search and clustering of dense vectors. For weak augmentations, we employ normalization, horizontal flipping and random cropping. For strong augmentations, we use RandAugment strategy for all, which selects the type and magnitude of augmentation based on uniform probability.

For implementation of SSL methods, we firstly pre-train the network with complete CLL dataset for 200 epochs. Next, we perform pseudo-labeling iteratively and train the model for another 200 epochs.

All of the experiments are implemented based on PyTorch (Paszke et al., 2019) and all of our experiments are conducted with 8 NVIDIA 4090 GPUs.

# G MORE EXPERIMENTAL RESULTS.

## G.1 QUANTATIVE RESULTS: PERFORMANCE OF PLG AND NLE.

As shown in Table 5, PLG achieves high selected ratio in all datasets, even reaching 98.86% on FMNIST (MCLL), and 76.08% on the difficult CIFAR-100. And the precision remains high even at a high selected ratio. NLE also has high precision, reaching 80.84% even on CIFAR-100.

## G.2 PARAMETER SENSITIVITY ANALYSIS.

**Influence of memory bank size $t$.** As is shown in Table 6, it is obvious that $t$ has little effect on the experimental results on the three datasets. We chose $t = 5$, which has a slightly better effect. It is important to utilize historical confidence information to alleviate confirmation bias, but the size of the memory banks does not matter.

Table 5: The performance of PLG and NLE on three settings. The results (mean $\pm$ std) are reported over 3 random trials. $\eta$ denotes PLG selected ratio, $1-\epsilon_1$ denotes PLG precision, $1-\epsilon_2$ denotes NLE precision.

| Dataset | Case | Performance | | |
|---|---|---|---|---|
| | | $\eta$ | 1-$\epsilon_1$ | 1-$\epsilon_2$ |
| CIFAR-10 | SCLL | 93.40±0.14% | 96.79±0.08% | 97.79±0.09% |
| | MCLL | 97.49±0.12% | 99.30±0.05% | 98.91±0.07% |
| FMNIST | SCLL | 97.25±0.08% | 95.24±0.11% | 97.56±0.08% |
| | MCLL | 98.86±0.06% | 97.89±0.11% | 98.97±0.03% |
| CIFAR-100 | MCLL | 76.08±0.22% | 79.84±0.41% | 80.84±0.04% |

Table 6: Classification accuracy of PLNL with different memory bank size $t$ on three benchmark datasets. The best results are highlighted in bold and the second best are underlined (The same applies hereinafter).

| t | STL-10 SCLL | CIFAR-10 SCLL | CIFAR-100 MCLL |
|---|---|---|---|
| 2 | 54.45% | 94.12% | <u>64.21%</u> |
| 3 | <u>54.95%</u> | 93.98% | 63.95% |
| 4 | 54.88% | <u>94.56%</u> | 63.81% |
| 5 | **55.25%** | **94.78%** | **64.33%** |

**Influence of confidence threshold $\lambda$.** From Table 8, we observe that there is a trade-off between $\eta$ and 1-$\epsilon_1$, i.e., a higher threshold will lead the precision to increase but result in less reliable instances selected while a lower threshold will decrease the precision but select more reliable instances. Instance-aware self-adaptive threshold (IST) shows obvious performance gain compared with a fixed global threshold, showcasing its effectiveness.

Table 7: Classification accuracy of PLNL with different $k$-NN parameter $k$ on three benchmark datasets.

| $k$ | STL-10 SCLL | CIFAR-10 SCLL | CIFAR-100 MCLL |
|---|---|---|---|
| 100 | <u>54.45%</u> | 94.25% | <u>63.27%</u> |
| 250 | **55.25%** | <u>94.73%</u> | **64.33%** |
| 500 | 47.23% | **94.78%** | 59.65% |

**Influence of $k$-NN parameter $k$.** As is shown in Table 7, the selection of $k$ depends on the specific situation of the dataset. For example, STL-10 has only 500 labeled instances for each category. If $k$ is set too large, there will be instances that are not in the category near the decision boundary, which will induce more noise in NLE and cause performance degradation.

Table 8: Classification accuracy and PLG performance of PLNL with different confidence threshold $\lambda$ on two benchmark datasets. $\eta$ denotes PLG selected ratio, $1-\epsilon_1$ denotes PLG precision, $1-\epsilon_2$ denotes NLE precision. IST denotes instance-aware self-adaptive threshold.

| $\lambda$ | CIFAR-10, SCLL | | | CIFAR-100, MCLL | | |
|---|---|---|---|---|---|---|
| | Accuracy | $\eta$ | 1-$\epsilon_1$ | Accuracy | $\eta$ | 1-$\epsilon_1$ |
| 0.85 | 92.89% | 99.86% | 90.79% | 62.12% | 81.52% | 72.88% |
| 0.90 | 93.85% | 95.78% | 92.43% | 63.95% | 77.25% | 77.34% |
| 0.95 | <u>94.74%</u> | 87.28% | 98.68% | <u>64.12%</u> | 73.48% | 83.61% |
| IST ($\alpha = 0.5$) | **94.78%** | 93.40% | 96.79% | **64.33%** | 76.08% | 79.84% |

### G.3 Experimental results on Tiny ImageNet

Tiny-ImageNet (Le and Yang, 2015) contains 100000 images of 200 classes. Each class has 500 training images, 50 validation images and 50 test images. Due to its huge number of categories, it is an extremely difficult dataset for complementary label learning. Most existing CLL methods have only tested their performance on 10-class small datasets. Most of their backbones are ResNet and

Table 9: Comparison of classification accuracies between different methods on Tiny-ImageNet with multiple complementary labels per instance.

| Method | Tiny-ImageNet |
|---|---|
| UB-EXP | 3.89% |
| UB-LOG | 7.17% |
| SCL-EXP | 3.36% |
| SCL-LOG | 8.96% |
| POCR | 4.29% |
| SELF-CL | 7.87% |
| ComCo | 8.52% |
| Ours | **11.87%** |

a single complementary label is a rather difficult setting for large datasets, these will lead to poor performance of traditional CLL methods.

However, we have tested the performance of our method on Tiny-ImageNet with MCLL settings. Though most of the methods perform poorly, our method still outperforms traditional CLL methods obtrusively.

### G.4 Extract features based on visual language models (VLMs) and self-supervised learning (SSL) techniques.

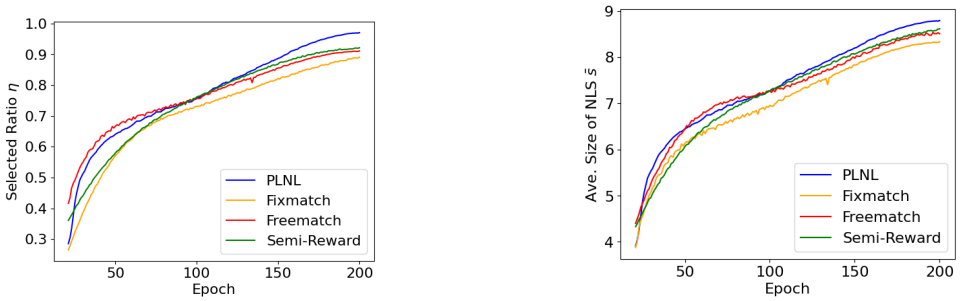
In the paper, we compute $k$-NN instances based on the model output space information, namely the feature extracted by the model itself. It strikes us that different features may have much to do with the NLE performance. To confirm this, we employ different feature extractors for computing $k$-NN, including PreActResNet-18-MoCo, BLIP-2 (Li et al., 2023a). Note that these results are just for performance comparison which has nothing to do with the results presented in the main body of the paper. For MoCo, we train a PreActResNet by self-supervised learning method MoCo (He et al., 2020) without any supervision. The weak and strong data augmentations used in MoCo follow the original configurations mentioned in the main body. Then we compute $k$-NN on the 512-dimensional feature output of the PreActResNet. For BLIP-2, we first employ the visual encoder to extract 768-dimensional high-quality representations and then leverage *faiss* (Johnson et al., 2019) to compute $k$-NN instances in this feature space. We compute the average precision of NLE $1-\epsilon_2$ and accuracy on CIFAR-10 (SCLL) and CIFAR-100 (MCLL). As shown in Table 10, the feature extracted from BLIP-2 outperforms MoCo and ResNet itself significantly. This shows the powerful visual representation ability of VLMs, which has a great potential for facilitating innovation in weakly-supervised learning in the future.

Table 10: Comparison of classification accuracies and NLE precision between different methods on CIFAR-100 with multiple complementary labels per instance. PreActResNet-18 denotes leveraging the model output space information for $k$-NN calculation

| Feature Extractor | CIFAR-10, SCLL | | CIFAR-100, MCLL | |
|---|---|---|---|---|
| | Accuracy | $1-\epsilon_2$ | Accuracy | $1-\epsilon_2$ |
| MoCo | 93.12% | 95.64% | 61.82% | 75.34% |
| BLIP-2 | **95.84%** | **99.91%** | **69.85%** | **93.34%** |
| PreActResNet-18 | 94.78% | 97.79% | 64.33% | 80.84% |

### G.5 Comparison with SemiReward

We further compare the performance of PLNL with one of the recently published SSL method SemiReward Li et al. (2023b). It can be seen that PLNL still outperforms SemiReward significantly in selection ratio and average size of NLS.




(a) Selected Ratio vs. Epoch

(b) Ave. Num. of NLS vs. Epoch

Figure 4: The experiments is conducted on CIFAR-10 with single complementary labels (SCLL). (a) shows that selected ratio of PLNL trenscends Fixmatch, Freematch and SemiReward significantly. (b) shows that average size of NLS of PLNL is significantly larger due to specially designed technique NLE for enhancing the untrustworthy negative labels. Nearly all negative labels are revealed at the end of training, almost reaching 9 negative labels for each instances in CIFAR-10.

## H DEVIATION REPORTS OF TABLE 3 AND TABLE 4

Table 11: The performance of PLNL with single network on two settings.

| Method | CIFAR-10 SCLL | | CIFAR-100 MCLL | |
|---|---|---|---|---|
| | $\eta$ | 1-$\epsilon_1$ | $\eta$ | 1-$\epsilon_1$ |
| Single | 84.65±0.12% | 90.21±0.09% | 67.43±0.24% | 70.62±0.31% |
| Two-view | **93.40±0.14%** | **96.79±0.08%** | **76.08±0.22%** | **79.84±0.41%** |

Table 12: Classification accuracy of degenerated methods on three settings.

| Method | STL-10 SCLL | CIFAR-10 SCLL | CIFAR-100 MCLL |
|---|---|---|---|
| PLNL | **55.25±0.35%** | **94.78±0.12%** | **64.33±0.43%** |
| PLNL v1 | 49.25±0.41% | 93.75±0.08% | 63.09±0.27% |
| PLNL v2 | 49.82±0.39% | 92.01±0.12% | 58.94±0.31% |
| PLNL v3 | 53.22±0.35% | 94.28±0.10% | 63.14±0.32% |
| POCR | 34.96±0.32% | 94.15±0.09% | 53.16±0.11% |

# I DETAILED DERIVATION OF EQUATION 30

$$\begin{aligned}\widehat{R}'(f) &= \frac{1}{N}\sum_{i=1}^{N}\bar{\mathcal{L}}_{CLL}(f(\boldsymbol{x}_i),\widehat{\bar{Y}}_i)\\ &= \frac{1}{N}\sum_{i=1}^{N}\sum\nolimits_{y\notin\widehat{\bar{Y}}_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}),y),\end{aligned} \tag{47}$$

This step is due to the definition of CLL loss function:

$$\mathcal{L}_{CLL}(f(\boldsymbol{x}),\widehat{\bar{Y}}_i)=\sum\nolimits_{y\notin\widehat{\bar{Y}}_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}),y)$$

Then we have:

$$\begin{aligned}\widehat{R}'(f) = &\frac{1}{N}\sum_{i=1}^{N}\ell(f(\boldsymbol{x}_i),y_i)+\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i\notin\widehat{\bar{Y}}_i)\left[\sum_{c\notin\widehat{\bar{Y}}_i,c\neq y_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),c)\underline{\underline{-\frac{K-|\widehat{\bar{Y}}_i|-1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),y_i)}}\right]\\ &+\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i\in\widehat{\bar{Y}}_i)\left[\sum_{c\notin\widehat{\bar{Y}}_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),c)\underline{\underline{-\ell(f(\boldsymbol{x}_i),y_i)}}\right]\end{aligned} \tag{48}$$

The first term is the loss on the ground-truth label, summing up to the empirical risk $\widehat{R}(f)$. The second and third term is the difference between empirical risk $\widehat{R}(f)$ and practical empirical risk with pseudo-labeling error $\widehat{R}'(f)$.

Since pseudo-labeling error may occur during PLG and NLE processes, the ground-truth label of $\boldsymbol{x}_i$ may be mistakenly included or correctly excluded in the enhanced negative label set $\widehat{\bar{Y}}_i$. We discuss two situations separately (i.e. $\mathbb{I}(y_i\in\widehat{\bar{Y}}_i)$ and $\mathbb{I}(y_i\notin\widehat{\bar{Y}}_i)$). We extract the empirical risk term $\widehat{R}(f)$ by subtracting the equivalent value in the second and third terms, which is shown between square brackets above.

As $\ell(f(\boldsymbol{x}),y)$ is bounded by a positive value $B$. We can scale up the second and third terms by directly removing the terms after minus sign, which is double underlined above, which will yield the following line.

$$\widehat{R}'(f)\leq\widehat{R}(f)+\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i\notin\widehat{\bar{Y}}_i)\sum_{c\notin\widehat{\bar{Y}}_i,c\neq y_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),c) \tag{49}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i\in\widehat{\bar{Y}}_i)\sum_{c\notin\widehat{\bar{Y}}_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),c), \tag{50}$$

Then we utilize Jensen's Inequality for further scaling up.

Jensen's Inequality for concave function $\varphi$:

$$\varphi\left(\sum_{i=1}^{n}g(x_i)\lambda_i\right)\geq\sum_{i=1}^{n}\varphi(g(x_i))\lambda_i,$$

where $\lambda_1+\lambda_2+\cdots+\lambda_n=1,\lambda_i\geq 0$.

Here, the concave function is $\varphi(|\widehat{\bar{Y}}_i|) = \frac{1}{K-|\widehat{\bar{Y}}_i|}$, where $\lambda_i = \frac{1}{N}$, then we have

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),c) \leq \sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i}\frac{1}{K-\frac{1}{N}\sum_{i=1}^{N}|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),c)$$

Since we have defined that $\bar{s} = \frac{1}{N}\sum_{i=1}^{N}|\widehat{\bar{Y}}_i|$ and $\mathbb{I}(y_i \notin \widehat{\bar{Y}}_i) = 1-\epsilon$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i \notin \widehat{\bar{Y}}_i)\sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i}\frac{1}{K-|\widehat{\bar{Y}}_i|}\ell(f(\boldsymbol{x}_i),c) &\leq (1-\epsilon)\frac{1}{K-\bar{s}}\sum_{c\notin\widehat{\bar{Y}}_i, c\neq y_i}\ell(f(\boldsymbol{x}_i),c),\\ &\leq (1-\epsilon)\frac{K-\bar{s}-1}{K-\bar{s}}B, \end{aligned} \tag{51}$$

The third term in Eq.(6) follows similar scaling procedures.