CLImage: Human-Annotated Datasets for Complementary-Label Learning

Anonymous authors
Paper under double-blind review

Abstract

Complementary-label learning (CLL) is a weakly-supervised learning paradigm that aims to train a multi-class classifier using only complementary labels, which indicate classes to which an instance does not belong. Despite numerous algorithmic proposals for CLL, their practical applicability remains unverified for two reasons. Firstly, these algorithms often rely on assumptions about the generation of complementary labels, and it is not clear how far the assumptions are from reality. Secondly, their evaluation has been limited to synthetic datasets. To gain insights into the real-world performance of CLL algorithms, we developed a protocol to collect complementary labels from human annotators. Our efforts resulted in the creation of four datasets: CLCIFAR10, CLCIFAR20, CLMicroImageNet10, and CLMicroImageNet20, derived from well-known classification datasets CIFAR10, CIFAR100, and TinyImageNet200. These datasets represent the very first real-world CLL datasets. Through extensive benchmark experiments, we discovered a notable decrease in performance when transitioning from synthetic datasets to real-world datasets. We investigated the key factors contributing to the decrease with a thorough dataset-level ablation study. Our analyses highlight annotation noise as the most influential factor in the real-world datasets. In addition, we discover that the biased-nature of human-annotated complementary labels and the difficulty to validate with only complementary labels are two outstanding barriers to practical CLL. These findings suggest that the community focus more research efforts on developing CLL algorithms and validation schemes that are robust to noisy and biased complementary-label distributions.

1 Introduction

Ordinary multi-class classification methods rely heavily on high-quality labels to train effective classifiers. However, such labels can be expensive and time-consuming to collect in many real-world applications. To address this challenge, researchers have turned their attention towards weakly-supervised learning, which aims to learn from incomplete, inexact, or inaccurate data sources (Zhou, 2018; Sugiyama et al., 2022). This learning paradigm includes but is not limited to noisy-label learning (Frénay & Verleysen, 2013), partial-label learning (Cour et al., 2011), positive-unlabeled learning (Denis, 1998), and complementary-label learning (Ishida et al., 2017).

In this work, we focus on complementary-label learning (CLL). This learning problem involves training a multi-class classifier using only complementary labels, which indicate the classes that a data instance does not belong to. Although several algorithms have been proposed to learn from complementary labels, they were only benchmarked on synthetic datasets with some idealistic assumptions on complementary-label generation (Ishida et al., 2017; 2019; Chou et al., 2020; Wang et al., 2021; Liu et al., 2023). Thus, it remains unclear how well these algorithms perform in practical scenarios.

In particular, current CLL algorithms heavily rely on the *uniform assumption* for generating complementary labels (Ishida et al., 2017), which specifies that complementary labels are generated by uniformly sampling from the set of all possible complementary labels. To alleviate the restrictiveness of the uniform assumption, (Yu et al., 2018) considered a more general *class-conditional assumption*, where the distribution of the complementary labels only depends on its ordinary labels. These assumptions have been used in many subsequent works to generate the *synthetic complementary datasets* for examining CLL algorithms (Ishida et al., 2019; Chou et al., 2020; Wang et al., 2021; Wei et al., 2022; Liu et al., 2023). Although these assumptions simplify the design and analysis of CLL algorithms, it remains unknown whether these assumptions hold true in practice and whether violation of these assumptions will significantly affect the performance of CLL algorithms. In addition to the uniform or class-conditional assumptions, most existing studies implicitly assumes that the complementary labels are noise-free. That is, they do not mistakenly represent the ordinary labels. While some studies claim to be more robust to noisy complementary labels (Lin & Lin, 2023), they were only tested on synthetic scenarios. It remains unclear how noisy the real-world datasets are, and how such noise affects the performance of current CLL algorithms.

To understand how much the real-world scenario differs from the assumptions, we collect human-annotated complementary-label datasets and conduct benchmarking experiments. We begin by constructing the CLCIFAR10 and CLCIFAR20 datasets, derived from the widely used CIFAR datasets for multi-class classification (Krizhevsky, 2012). Building upon this foundation, we further extend our collection to include two additional human-annotated datasets, CLMicroImageNet10 and CLMicroImageNet20, derived from TinyImageNet200 (Le & Yang, 2015). For all four datasets, we analyze the collected complementary labels, including their noise rates and non-uniform nature. Then, we perform benchmark experiments with diverse state-of-the-art CLL algorithms and conduct dataset-level ablation study on the assumptions of complementary-label generation using the collected datasets.

Our studies reveal annotation noise as the most influential factor affecting the performance of CLL algorithms in real-world datasets. The claim is evidenced by our ablation study results on both synthetic and real-world complementary datasets. Notably, classification accuracy drops from 64.18% on CIFAR10 to 34.85% on CLCIFAR10, highlighting the detrimental impact of noisy complementary labels. To further investigate the role of label noise in the performance gap across different CLL algorithms, we conducted an ablation study that examined various noise levels. The results consistently reinforce our conclusion that the performance disparity between human-annotated complementary labels and synthetically generated complementary labels is primarily driven by label noise. Moreover, through thorough analysis we confirmed that the non-uniform nature of human-annotated complementary labels makes certain CLL algorithms more susceptible to overfitting, although its impact is less pronounced compared to label noise.

These findings immediately suggest that the community focus more research efforts on developing CLL algorithms that are robust to noisy and non-uniform complementary-label distributions. In addition, we used the collected datasets to demonstrate that existing complementary-label-only validation schemes are not mature yet, suggesting the community a novel research direction for making CLL practical. Our contributions are summarized as follows:

- We designed a collection protocol of complementary labels (CLs) for images, and verified that the protocol collects reasonable human-annotated CLs across different datasets.
- We collected the set of four real-world CL datasets and plan to release **CLImage** to support the continuous research of the community.
- We analyzed the collected datasets with extensive benchmarking experiments, which provides novel and valuable insights for the community.

2 Preliminaries on CLL

2.1 Complementary-label learning

In ordinary multi-class classification, a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ that is *i.i.d.* sampled from an unknown distribution is given to the learning algorithm. For each i, $\mathbf{x}_i \in \mathbb{R}^M$ represents the M-dimension feature of the i-th instance and $y_i \in [K] = \{1, 2, \ldots, K\}$ represents the class $\mathbf{x}_i$ belongs to. The goal of the learning algorithm is to learn a classifier from D that can predict the labels of unseen instances correctly. The classifier is typically parameterized by a scoring function $\mathbf{g} \colon \mathbb{R}^M \to \mathbb{R}^K$, and the prediction is made by

$\arg\max_{k\in[K]} \mathbf{g}(\mathbf{x})_k$ given an instance $\mathbf{x}$, where $\mathbf{g}(\mathbf{x})_k$ denotes the k-th output of $\mathbf{g}(\mathbf{x})$. In contrast to ordinary multi-class classification, CLL shares the same goal of learning a classifier but trains with different labels. In CLL, the ordinary label y_i is not accessible to the learning algorithm. Instead, a complementary label $\bar{y}_i$ is provided, which is a class that the instance $\mathbf{x}_i$ does *not* belong to. The goal of CLL is to learn a classifier that is able to predict the correct labels of unseen instances from a complementary-label dataset $\bar{D} = \{(\mathbf{x}_i, \bar{y}_i)\}_{i=1}^n$.

2.2 Common assumptions on CLL

Researchers have made some additional assumptions on the generation process of complementary labels to facilitate the analysis and design of CLL algorithms. One common assumption is the *class-conditional assumption* (Yu et al., 2018). It assumes that the distribution of a complementary label only depends on its ordinary label and is independent of the underlying example's feature, i.e., $P(\bar{y}_i \mid \mathbf{x}_i, y_i) = P(\bar{y}_i \mid y_i)$ for each i. One special case of the class-conditional assumption is the *uniform assumption*, which further specifies that the complementary labels are generated uniformly. That is, $P(\bar{y}_i = k|y_i = j) = \frac{1}{K-1}$ for all $k \in [K]\backslash\{j\}$ (Ishida et al., 2017; 2019; Lin & Lin, 2023).

For convenience, a $K \times K$ matrix T, called *transition matrix*, is often used to represent how the complementary labels are generated under the class-conditional assumption. $T_{j,k}$ is defined to be the probability of obtaining a complementary label k if the underlying ordinary label is j, i.e., $T_{j,k} = P(\bar{y} = k \mid y = j)$ for each $j, k \in [K]$. The diagonals of T hold the conditional probabilities that a complementary label mistakenly represents the ordinary label. That is, they indicate the noise level of the complementary labels. When T contains all zeros on its diagonals, the CLL scenario is called *noiseless.* For instance, the uniform and noiseless assumption can be represented by $T_{j,j} = 0$ for each $j \in [K]$ and $T_{j,k} = \frac{1}{K-1}$ for each $k \neq j$. Class-conditional CLL scenarios based on any other non-uniform T are often called *biased.*

2.3 A brief overview of CLL algorithms

The pioneering work by Ishida et al. (2017) studied how to learn from complementary labels under the *uniform assumption* by converting the risk estimator in ordinary multi-class classification to an unbiased risk estimator (**URE**) in CLL (Ishida et al., 2017). **URE** is then found to be prone to overfitting because of negative empirical risks, and is upgraded with two tricks, non-negative risk estimator (**URE-NN**) and gradient accent (**URE-GA**) (Ishida et al., 2019). The *surrogate complementary loss* (**SCL**) algorithm mitigates the overfitting issue of **URE** by a different loss design that decreases the variance of the empirical estimation. However, these algorithms either rely on the uniform assumption in design or are only tested on the synthetic datasets that obeys the uniform assumption.

To make CLL one step closer to practice, researchers have explored algorithms to go beyond the uniform (and thus noiseless) assumption. (Yu et al., 2018) utilized the forward-correction loss (**FWD**) to accommodate biased complementary label generation by adapting techniques from noisy label learning (Patrini et al., 2017) to change the loss. Additionally, (Gao & Zhang, 2021) proposed the **L-W** algorithm based on discriminatively modeling the distribution of complementary labels through a weighting function, further improving the performance in bias scenario. Furthermore, (Ishiguro et al., 2022) designed robust loss functions for learning from noisy CLs, including **MAE** and **WMAE**, by applying the gradient ascent technique (Ishida et al., 2019) to handle noisy scenarios.

Besides CLL algorithms, a crucial component for making CLL practical is model validation. In ordinary-label learning, this can be done by naively calculating the classification accuracy on a validation dataset. In CLL, this scheme can be intractable if there are not enough ordinary labels. One generic way of model validation is based on the result of (Ishida et al., 2019) by calculating the unbiased risk estimator of the zero-one loss, i.e.,

$$\hat{R}_{01}(\mathbf{g}) = \frac{1}{N} \sum_{i=1}^{N} e_{\bar{y}_i}^{\top} (T^{-1}) \ell_{01}(\mathbf{g}(x_i)) \tag{1}$$

where $e_{\overline{y}_i}$ denotes the one-hot vector of $\overline{y}_i$, $\ell_{01}(\mathbf{g}(x_i))$ denotes the K-dimensional vector $(\ell_{01}(\mathbf{g}(x_i), 1), \ldots, \ell_{01}(\mathbf{g}(x_i)), K))^T$, and $\ell_{01}(\mathbf{g}(x_i), k) = 0$ if $\arg\max_{k\in[K]} \mathbf{g}(x_i) = k$ and 1 otherwise, representing the zero-one loss of $\mathbf{g}(x_i)$ if the ordinary label is k. This estimator will be used in the experiments in Section 6. Another validation objective, surrogate complementary esimation loss (SCEL), was proposed by (Lin & Lin, 2023). SCEL measures the log loss of the complementary probability estimates induced by the probability estimates on the ordinary label space. The formula to calculate SCEL is as follows,

$$\hat{R}_{\text{SCEL}}(\mathbf{g}) = \frac{1}{N} \sum_{i=1}^{N} -\log \left(e_{\bar{y}_i}^{\top} T^{\top} \operatorname{softmax}(\mathbf{g}(x_i)) \right). \tag{2}$$

3 Construction of the CLImage collection

In this section, we introduce the four complementary-labeled datasets that we collected, CLCIFAR10, CLCIFAR20, CLMicroImageNet10 and CLMicroImageNet20. All datasets are labeled by human annotators on Amazon Mechanical Turk (MTurk)[1].

3.1 Datasets and goals

The complementary-labeled datasets are derived from ordinary multi-class classification datasets. CIFAR10, CIFAR100 and TinyImageNet200 (Krizhevsky, 2012; Le & Yang, 2015; Russakovsky et al., 2015). This selection is motivated by the real-world noisy label dataset by (Wei et al., 2022). Building upon the CIFAR and TinyImageNet200 datasets allow us to estimate the noise rate and the empirical transition matrix easily, as they already contain nearly noise-free ordinary labels. In addition, many of the state-of-the-art CLL algorithms have been benchmarked on synthetic complementary labels with the CIFAR datasets (Dosovitskiy et al., 2020; Kolesnikov et al., 2020; Oquab et al., 2023). Our CLCIFAR counterparts immediately allow a fair comparison to those results with the same network architecture.

In addition to our CLCIFAR extensions, we are the first to introduce (Tiny)ImageNet-derived datasets to the CLL literature. Such datasets serve two purposes. First, it allows us to confirm the validity of our collection protocol and findings beyond CIFAR-derived datasets. Second, ImageNet knowingly contains images of higher complexity than CIFAR and can thus be used to challenge the ability of existing CLL algorithms more realistically.

There is a historical note that is worth sharing with the community: We initially attempted to collect complementary labels based on the 100 classes in CIFAR100. But some preliminary testing soon revealed that state-of-the-art CLL algorithms cannot produce meaningful classifiers for 100 classes even on synthetic complementary labels that are uniformly and noiselessly generated. We thus set our collection goals to be 10-class classification, which is the focus of most current CLL studies, and 20-class classification, which extends the horizon of CLL and matches the 20 super-class structure in CIFAR.

3.2 Complementary label collection protocol

To collect only complementary labels from the CIFAR, TinyImageNet datasets, for each image in the training split, we first randomly sample four distinct labels and ask the human annotators to select any of the *incorrect* one from them. To leave room for analyzing the annotators' behavior, each image is labeled by three different annotators. The four labels are re-sampled for each annotator on each image. That is, each annotator possibly receives a different set of four labels to choose from. An algorithmic description of the protocol is as follows. For each image $\mathbf{x}$,

1. Uniformly sample four labels without replacement from the label set $[K]$.
2. Ask the annotator to select any one of the complementary label $\bar{y}$ from the four sampled labels.
3. Add the pair $(\mathbf{x}, \bar{y})$ to the complementary dataset.

Note that if the annotators always select one of the correct complementary labels uniformly, the empirical transition matrix will also be uniform in expectation. We will inspect the empirical transition matrix in

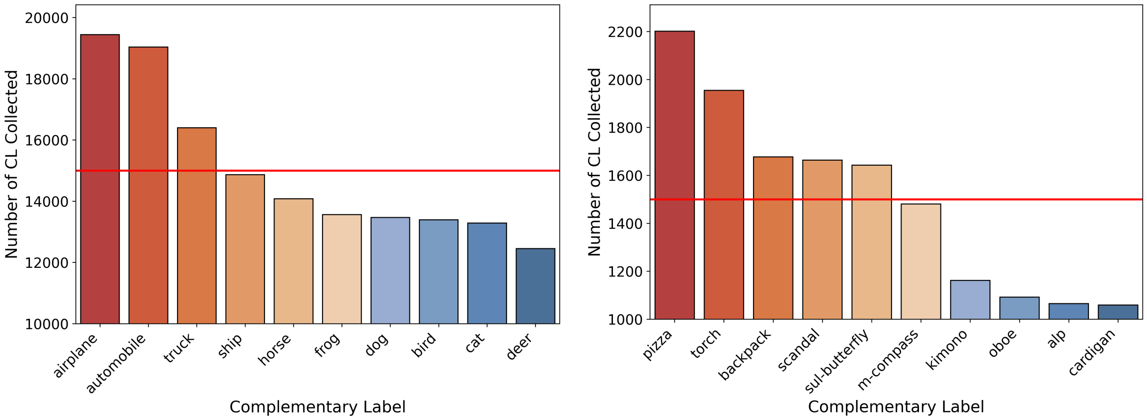
---

[1] `https://www.mturk.com/`

Section 4. The labeling tasks are deployed on MTurk by dividing them into smaller we first divide the total images into smaller human intelligence tasks (HITs). For instance, for constructing the CLCIFAR datasets, we first divide the 50,000 images into five batches of 10,000 images. Then, each batch is further divided into 1,000 HITs with each HIT containing 10 images. Each HIT is deployed to three annotators, who receive 0.03 dollar as the reward by annotating 10 images. To make the labeling task easier and increase clarity, the size of the images are enlarged to 200 × 200 pixels.

4 Dataset Characteristic

Next, we closely examine the collected complementary labels. We first analyze the error rates of the collected labels, and then verify whether the transition matrix is uniform or not. Finally, we end with an analysis on the behavior of the human annotators observed in the label collection protocol.



(a) CLCIFAR10

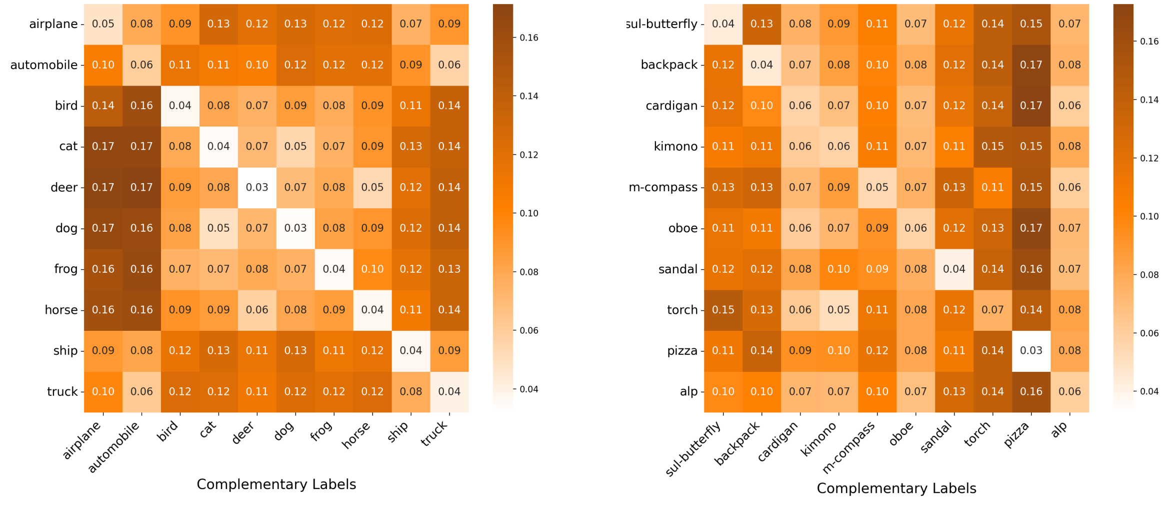
(b) CLMicroImageNet10

Figure 1: The label distribution of CLCIFAR10 and CLMicroImageNet10 datasets.

Observation 1: noise rate compared to ordinary label collection We first look at the noise rate of the collected complementary labels. A complementary label is considered to be incorrect if it is actually the ordinary label. The mean error rate made by the human annotators is 3.93% for CLCIFAR10, 2.80% for CLCIFAR20, 5.19% for CLMicroImageNet10 and 3.21% for CLMicroImageNet20. In theory, we can estimate a random annotator achieves a noise rate of $\frac{1}{K}$ for complementary label annotation and a noise rate of $\frac{K-1}{K}$ for ordinary label annotation. If we compare the human annotators to a random annotator, then for CLCIFAR10, human annotators have 60.7% less noisy labels than the random annotator whereas for CIFAR10-N, human anotators have 78.17% less noisy labels. This demonstrates that human annotators are more competent compared to a random annotator in the ordinary-label annotation. Similarly, human annotators have 44% less noise than a random annotator for CLCIFAR20 and 73.05% less noise for CIFAR100N-coarse (Wei et al., 2022). This observation reveals that while the absolute noise rate is lower in annotating complementary labels, it may be more difficult to be competent against random labels than the ordinary label annotation.

Observation 2: imbalanced complementary label annotation Next, we analyze the distribution of the collected complementary labels. The frequency of the complementary labels for the CLCIFAR10 and CLMicroImageNet10 (CLMIN10) datasets are reported in Figure 1. As we can see in the figure, the annotators exhibit specific biases towards certain labels. For instance, in CLCIFAR10, annotators prefer “airplane” and “automobile”, while in CLMIN10, they prefer “pizza” and “torch”. In CLCIFAR10, the bias is towards labels in different categories, as vehicles (“airplane”, “automobile”) versus animals (“cat”, “bird”). In contrast, in CLMIN10, the bias is towards items that are easily recognizable (“pizza” and “torch”) and against those that are less familiar (“cardigan” or “alp”).

Observation 3: biased transition matrix Finally, we visualize the empirical transition matrix using the collected CLs in Figure 2. Based on the first two observations, we could imagine that the transition


(a) CLCIFAR10

(b) CLMicroImageNet10

Figure 2: The empirical transition matrices of CLCIFAR10 and CLMicroImageNet10.

matrix is biased. By inspecting Figure 2, we further discover that the bias in the complementary labels are dependent on the true labels. For instance, in CLCIFAR10, despite we see more annotations on airplane and automobile in aggregate, conditioning on the transportation-related labels (“airplane”, “automobile”, etc), the distribution of the complementary labels becomes more biased towards other animal-related labels (“bird”, “cat”, etc.) Furthermore, this observation holds true on CLMIN10 as well. Next, we study the impact of the bias and noise on existing CLL algorithms.

We discovered similar patterns in all four human-annotated datasets, validating that our design methodology is practical for collecting real-world CLL image datasets. Due to space limitations, we have included the detailed analysis of CLCIFAR20 and CLMicroImageNet20 in Appendix A.4.

5 Experiments

In this section, we benchmarked several state-of-the-art CLL algorithms on CLImage. A significant performance gap between the models trained on the humanly annotated CLCIFAR, CLMicroImageNet dataset and those trained on the synthetically generated complementary labels (CL) was observed in Section 5.1, which motivates us to analyze the possible reasons for the gap with the following experiments. To do so, we discuss the effect of three factors in the label generating process, feature dependency, noise, and biasedness, in Section 5.2, Section 5.3, and Section 5.4, respectively. From our experiment results, we conclude that noise is the dominant factor affecting the performance of the CLL algorithms on CLCIFAR[2].

5.1 Standard benchmark on CLImage

Baseline methods Several state-of-the-art CLL algorithms were selected for this benchmark. Some of them take the transition matrix T as inputs, which we call T**-informed** methods, including two version of forward correction (Yu et al., 2018): **FWD-U** and **FWD-R**, two version of unbiased risk estimator with gradient ascent (Ishida et al., 2019): **URE-GA-U** and **URE-GA-R**, and robust loss (Ishiguro et al., 2022) for learning from noisy CL: **CCE**, **MAE**, **WMAE**, **GCE**, and **SL**[3]. We also included some algorithms that assume the transition matrix T to be uniform, called T**-agnostic** methods, including surrogate complementary

---

[2] Due to space and time constraints, we only provide the results and discussion on the CLCIFAR datasets.

[3] Due to space limitations, we only provided the results of MAE. The remaining results and discussions related to the robust loss methods can be found in Appendix A.3

Table 1: Standard benchmark results on CLCIFAR/ CLMicroImageNet(CLMIN) and uniform-CIFAR/ MicroImageNet(MIN) datasets. Mean accuracy (± standard deviation) on the testing dataset from four trials with different random seeds. Highest accuracy in each column is highlighted in bold.

	uniform-CIFAR10	CLCIFAR10	uniform-CIFAR20	CLCIFAR20	uniform-MIN10	CLMIN10	uniform-MIN20	CLMIN20
FWD-U	**64.19±0.57**	34.83±0.50	21.54±0.37	8.03±0.74	36.30±1.12	23.85±2.76	12.57±2.94	6.33±1.04
FWD-R	61.32±0.90	**38.13±0.88**	21.50±0.38	**20.27±0.53**	35.70±1.19	**30.15±1.83**	**14.85±1.75**	**10.60±0.82**
URE-GA-U	50.24±1.11	34.72±0.40	16.67±1.35	10.49±0.52	35.70±1.97	22.90±2.97	11.65±1.90	5.75±0.43
URE-GA-R	50.73±1.83	30.23±0.70	17.57±0.61	6.17±0.82	33.65±1.40	13.25±5.11	9.78±3.88	6.50±0.35
SCL-NL	63.76±0.09	34.77±0.60	21.37±1.18	8.02±0.36	**37.05±1.40**	21.80±1.85	13.00±2.80	6.17±0.49
SCL-EXP	63.29±1.02	35.18±0.67	**21.57±1.13**	7.70±0.41	36.55±1.28	24.80±1.14	12.95±3.38	5.58±0.13
L-W	54.32±0.41	32.99±1.01	19.59±0.99	7.71±0.35	33.80±2.66	23.80±2.64	12.70±2.35	6.40±0.29
L-UW	57.52±0.59	34.69±0.32	20.71±0.92	8.15±0.30	35.10±2.74	22.40±1.67	12.12±3.13	6.35±0.86
PC-sigmoid	37.78±0.80	32.15±0.80	14.48±0.47	12.11±0.46	29.10±0.98	23.15±0.46	10.72±1.38	6.90±1.04
ROB-MAE	59.38±0.63	20.23±1.02	18.17±1.31	5.40±0.59	31.50±1.81	14.15±0.68	6.35±0.86	5.38±0.33

	CIFAR10	CIFAR20	MIN10	MIN20
standard supervision	82.80±0.28	63.80±0.49	68.70±1.53	63.90±1.00

loss **SCL-NL** and **SCL-EXP** (Chou et al., 2020), discriminative modeling **L-W** and its weighted variant (**L-UW**) (Gao & Zhang, 2021), and pairwise-comparison (**PC**) with the sigmoid loss (Ishida et al., 2017). The details of the algorithms mentioned above are discussed in Appendix B.

Implementation details We collected and released three CLs per image to prepare for future studies. However, for this standard benchmark, we chose the first CL from the collected labels for each data instances to form a single CLL dataset, ensuring reproducibility. Then, we trained a ResNet18 (He et al., 2016) model using the baseline methods mentioned above on the single CLL dataset using the Adam optimizer for 300 epochs without learning rate scheduling. Detailed results from the ablation study on various neural network architectures, which further justify our choice of ResNet18 as the backbone, are available in Appendix A.6. The training settings included a fixed weight decay of 10^{-4} and a batch size of 512. The experiments were run with Tesla V100-SXM2. For better generalization, we applied standard data augmentation technique, `RandomHorizontalFlip`, `RandomCrop`, and normalization to each image. The learning rate was selected from $\{10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}, 10^{-5}\}$ using a 10% hold-out validation set. We selected the learning rate with the best classification accuracy on the validation dataset. Note that here we assumed the ordinary labels in the validation dataset are known. We will discuss other validation objectives that rely only on complementary labels in Section 6. As CLL algorithms are prone to overfitting (Ishida et al., 2019; Chou et al., 2020), some previous works did not use the model after training for evaluation. Instead, previous works were performed by evaluating the model on the validation dataset and selecting the epoch with the highest validation accuracy. In this work, we also follow the same aforementioned technique to validate testing set. For reference, we also performed the experiments on synthetically-generated CLL dataset, where the CLs were generated uniformly and noiselessly, denoted uniform-CIFAR.

Results and discussion As we can observe in Table 1, there is a significant performance gap between the humanly annotated dataset, CLCIFAR, and the synthetically generated dataset, uniform-CIFAR. The difference between the two datasets can be divided into three parts: (a) whether the generation of complementary labels depends on the feature, (b) whether there is noise, and (c) whether the complementary labels are generated with bias. A negative answer to those questions simplify the problem of CLL. We can gradually simplify CLCIFAR to uniform-CIFAR by chaining those assumptions as follows [4]:

$$\boxed{\text{CLCIFAR}} \xRightarrow[\text{Remove feature dependency}]{\text{Section 5.2}} \xRightarrow[\text{Remove noise}]{\text{Section 5.3}} \xRightarrow[\text{Remove biasedness}]{\text{Section 5.4}} \boxed{\text{uniform-CIFAR}}$$

---

[3] Note that FWD-R and URE-GA-R assume the empirical transition matrix T_e to be provided. The empirical transition matrix is computed from the labels in the training set, so it is slightly different from a uniform transition matrix T_u in the uniform-CIFAR datasets. As a result, the performances of FWD-R and URE-GA-R do not exactly match those of FWD-U and URE-GA-U, respectively, in the uniform-CIFAR datasets.

[4] The "interpolation" between CLCIFAR and uniform-CIFAR does not necessarily have to be this way. For instance, one can remove the biasedness before removing the noise. We chose this order to reflect the advance of CLL algorithms. First, researchers address the uniform case (Ishida et al., 2017), then generalize to the biased case (Yu et al., 2018), then consider noisy labels (Ishiguro et al., 2022). There is no work considering feature-dependent complementary labels yet.

In the following subsections, we will analyze how these three factors affect the performance of the CLL algorithms.

5.2 Feature dependency

In this experiment, we verified whether the performance gap resulted from the feature-dependent generation of practical CLs. Conceivably, even if two images belong to the same class, the distribution on the complementary labels could be different. On the other hand, the distributional difference could also be too small to affect model performance, e.g., if $P(\bar{y} \mid y, \mathbf{x}) \approx P(\bar{y} \mid y)$ for most $\mathbf{x}$. Consequently, we decided to further look into whether this assumption can explain the performance gap. To observe the effects of approximating $P(\bar{y} \mid y, \mathbf{x})$ with $P(\bar{y} \mid y)$, we generated two synthetic complementary datasets, CLCIFAR10-*iid* and CLCIFAR20-*iid* by i.i.d. sampling CLs from the empirical transition matrix in CLCIFAR10 and CLCIFAR20, respectively. We proceeded to benchmark the CLL algorithms on CLCIFAR-*iid* and presented the accuracy difference compared to CLCIFAR in Table 2.

Results and discussion From Table 2, we observed that the accuracy barely changes on the resampled CLCIFAR-*iid*, suggesting that even if the complementary labels in CLCIFAR could be feature-dependent, this dependency does not affect the model performance significantly. Hence, there might be other factors contributing to the performance gap.

Table 2: Mean accuracy difference (± standard deviation) of different CLL algorithms. A plus indicates the performance on is calculated as CLCIFAR-*i.i.d.* accuracy minus CLCIFAR accuracy.

	FWD-U	FWD-R	URE-GA-U	URE-GA-R	SCL-NL	SCL-EXP	L-W	L-UW	PC-sigmoid
CLCIFAR10-iid	-1.1$\pm$2.17	-0.36$\pm$1.15	-3.03$\pm$1.25	0.74$\pm$0.35	-0.67$\pm$1.81	-1.97$\pm$1.16	-2.5$\pm$0.56	-3.53$\pm$1.36	-2.03$\pm$2.05
CLCIFAR20-iid	-0.64$\pm$0.39	-3.53$\pm$1.13	-0.37$\pm$0.51	1.79$\pm$2.34	-0.28$\pm$0.61	-0.39$\pm$0.69	-0.5$\pm$1.37	-0.82$\pm$0.04	-2.24$\pm$0.52

5.3 Labeling noise

In this experiment, we further investigated the impact of the label noise on the performance gap. Specifically, we measured the accuracy on the noise-removed versions of CLCIFAR datasets, where varying percentages (0%, 25%, 50%, 75%, or 100%) of noisy labels are eliminated.

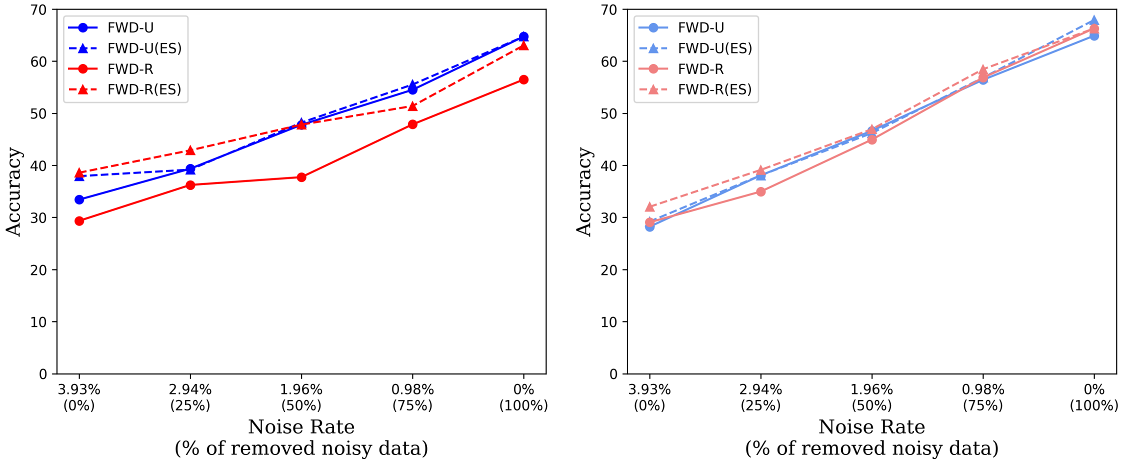
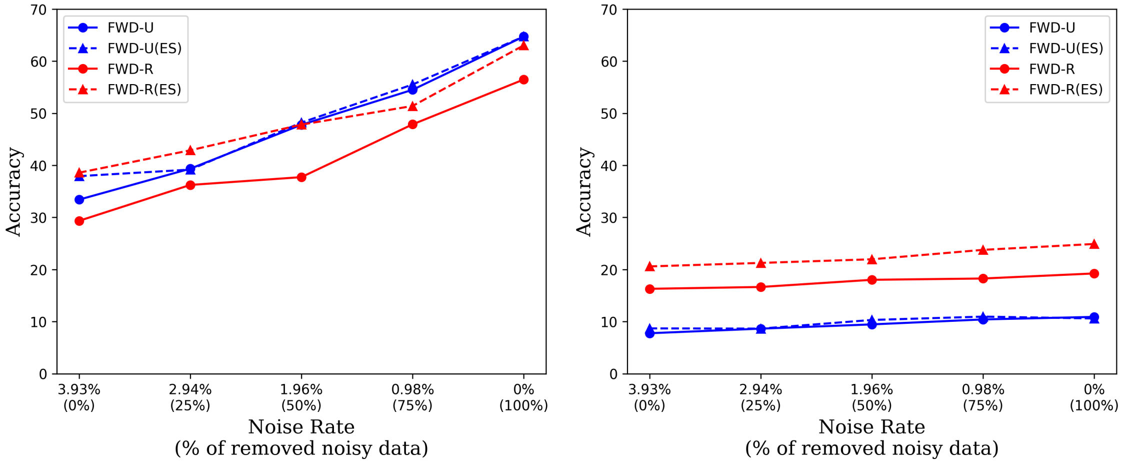
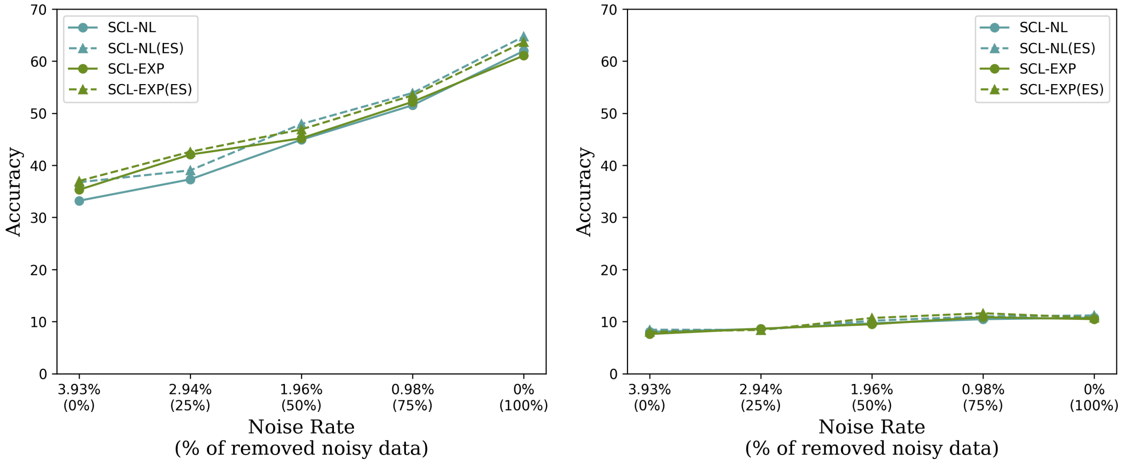


Figure 3: Accuracy of FWD-U and FWD-R on the noise-removed CLCIFAR10 dataset (**Left**) and the uniform-CIFAR10 dataset with uniform noise (**Right**) at varying noise rates.

Results and discussion We present the performance of FWD trained on the noise-removed CLCIFAR10 dataset in the left figure in Figure 3. From the figure, we observe a strong positive correlation between the performance and the proportion of removed noisy labels. When more noisy labels are removed, the performance gap diminishes and the accuracy approaches that of the ideal uniform-CIFAR dataset. Therefore, we conclude that the performance gap between the humanly annotated CLs and the synthetically generated CLs are primarily attributed to the label noise. The results for FWD-(U/R) and SCL-(NL/EXP) of the noise-removed CLCIFAR10 and CLCIFAR20 datasets are presented in the Figure 4. The other results for other algorithms can be found in Appendix C.



(a) FWD-(U/R) on CLCIFAR10 (b) FWD-(U/R) on CLCIFAR20



(c) SCL-(NL/EXP) on CLCIFAR10 (d) SCL-(NL/EXP) on CLCIFAR20

Figure 4: Accuracy of FWD-(U/R) and SCL-(NL/EXP) on the noise-removed CLCIFAR10 dataset (**Left**) and the CLCIFAR20 dataset with uniform noise (**Right**) at varying noise rates.

5.4 Biasedness of complementary labels

To further study the biasedness of CL as a potential factor contributing to the performance gap, we removed the biasedness from the noise-removed CLCIFAR dataset and examined the resulting accuracy. Specifically, we introduced the same level of uniform noise in uniform-CIFAR dataset and reevaluated the performance of FWD algorithms.

Results and discussion The striking similarity between the two curves in the right figure in Figure 3 shows that the accuracy is significantly influenced by label noise, while the biasedness of CL has a negligible impact on the results. Furthermore, we observe that the accuracy difference between the results of the last epoch and the best accuracy of validation set (or early-stopping: **ES**) results becomes smaller when the model is trained on the uniformly generated CLs. That is, the T-informed methods are more prone to overfitting when there is a bias in the CL generation.

With the experiment results in Section 5.2, 5.3, and 5.4, we can conclude that the performance gap between humanly annotated CL and synthetically generated CL is primarily attributed to label noise. Additionally, the biasedness of CLs may potentially contribute to overfitting, while the feature-dependent CLs do not detrimentally affect performance empirically. It is worth noting that in the last row of Table 1, the MAE methods that can learn from noisy CL fails to generalize well in the practical dataset. These results suggest that more research on learning with noisy complementary labels can potentially make CLL more realistic.

Following above conclusion, the label noise and biasedness of CL emerge as the two primary factors contributing to overfitting. To gain a better understanding, we conducted deeper investigation into this phenomenon. We demonstrated the necessity of employing data augmentation techniques to prevent overfitting and attempted to address the issue of overfitting by employing an interpolated transition matrix for regularization.

Ablation on data augmentation To further investigate the significance of data augmentation, we conducted identical experiments without employing data augmentation during the training phase. As we can observe in the training curves in Figure 5, data augmentation could improve the testing accuracy of all the algorithms we considered.

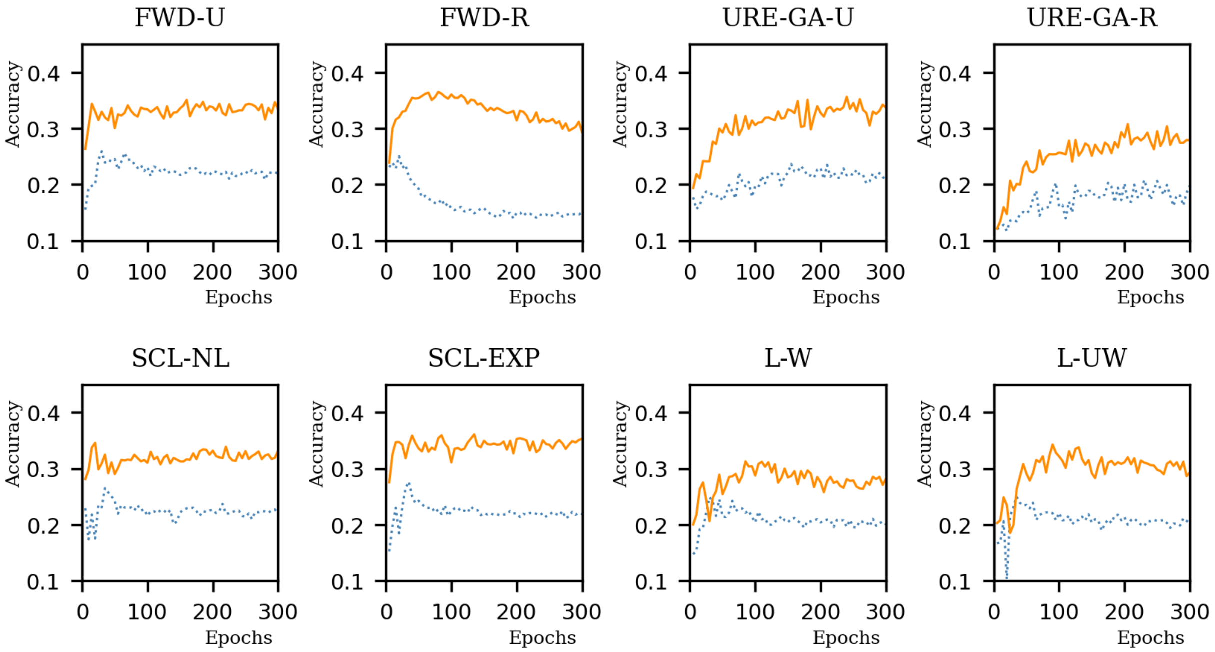


Figure 5: The Overfitting accuracy curve of FWD, URE, SCL-NL, L-W. The dotted line represents the accuracy obtained without data augmentation, while the solid line represents the accuracy with data augmentation included for reference. The accuracy of FWD, SCL-NL, SCL-EXP, L-W, L-UW methods reaches its highest at approximately the 50 epoches and converges to some lower point. The detail numbers are in Table 3

Ablation on interpolation between T_u and T_e In Table 1, we discovered that the T-informed methods did not always deliver better testing accuracy when T_e is given. Looking at the difference between the accuracy of using early-stopping and not using early-stopping, we observe that when the T_u is given to the T-informed methods, the difference becomes smaller. This suggests that T-informed methods using the empirical transition matrix has greater tendency to overfitting. On the other hand, T-informed methods using the uniform transition matrix could be a more robust choice.

Table 3: The overfitting results when there is no data augmentation.

	uniform-CIFAR10		CLCIFAR10		uniform-CIFAR20		CLCIFAR20	
methods	valid_acc	valid_acc (ES)	valid_acc	valid_acc (ES)	valid_acc	valid_acc (ES)	valid_acc	valid_acc (ES)
FWD-U	**48.44**	**49.33**	21.29	25.59	**17.4**	**17.97**	6.91	7.32
FWD-R	-	-	14.97	**28.3**	-	-	6.82	**14.67**
URE-GA-U	39.55	39.67	21.0	23.53	13.52	14.08	5.55	8.38
URE-GA-R	-	-	19.81	20.8	-	-	5.0	6.43
SCL-NL	48.2	48.27	**21.96**	26.51	16.55	17.54	**7.1**	7.92
SCL-EXP	46.79	47.52	21.89	27.66	16.18	17.89	6.9	7.3
L-W	27.02	44.78	20.06	27.6	10.39	16.3	5.64	8.02
L-UW	31.3	46.38	20.28	26.26	12.33	16.32	6.03	8.14
PC-sigmoid	18.97	33.26	-	-	7.67	10.41	-	-

We observe that the uniform transition matrix T_u acts like a regularization choice when the algorithms overfit on CLCIFAR. This results motivate us to study whether we can interpolate between T_u and T_e to let the algorithms utilize the information of transition matrix while preventing overfitting. To do so, we provide an interpolated transition matrix $T_{\text{int}} = \alpha T_u + (1-\alpha)T_e$ to the algorithm, where α controls the scale of the interpolation. As FWD is the T-informed method with the most sever overfitting when using T_u, we performed this experiment using FWD adn reported the results in Figure 6. As shown in Figure 6, FWD can learn better from an interpolated T_{int}, confirming the conjecture that T_u can serve as a regularization role.

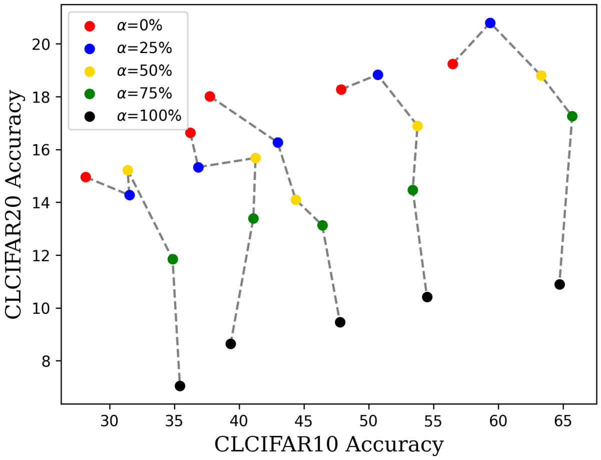


Figure 6: The last epoch accuracy of CLCIFAR10 and CLCIFAR20 for FWD algorithm with an α-interpolated transition matrix T_{int}. The five solid points on each cruve represent different noise cleaning rate: 0%, 25%, 50%, 75%, 100% from left to right.

6 Validation Objectives

Validation is a crucial component in applying CLL algorithms in practice. With the collection of the real-world datasets, we are now able to estimate the difference between using ordinary labels for validation (the common practice in existing CLL studies, as what we do in Section 5) and using complementary labels for validation.

Validation objectives As discussed in Section 2, validating the model performance solely with complementary labels poses a non-trivial challenge. To the best of our knowledge, only two existing CLL studies offer some possibility to evaluate a classifier *with only complementary labels*. They are URE (Ishida et al., 2019) and SCEL (Lin & Lin, 2023). We take these two validation objectives to select the optimal learning rate from $\{10^{-3},\ 5\times10^{-4},\ 10^{-4},\ 5\times10^{-5},\ 10^{-5}\}$ and provides the accuracy on testing set in Table 5. We compare the result to another validation objective that computes the accuracy on an equal number of *ordinary labels*. Our goal was to determine the gap between using complementary labels and ordinary labels for validation.

We selected the best learning rate based on the validation objectives for URE, SCEL, and ordinary-label accuracy, and then report the test performance, as shown in Table 4 for synthetic datasets and Table 5 for real-world datasets.

Table 4: The testing accuracy of models evaluated with URE and SCEL.

	uniform-CIFAR10				uniform-CIFAR20				uniform-MIN10				uniform-MIN20			
	URE	SCEL	valid acc	gap (↓)	URE	SCEL	valid acc	gap (↓)	URE	SCEL	valid acc	gap (↓)	URE	SCEL	valid acc	gap (↓)
FWD-U	$53.41_{\pm 5.51}$	$50.36_{\pm 3.25}$	$64.19_{\pm 0.57}$	10.78	$16.73_{\pm 2.29}$	$16.52_{\pm 2.61}$	$21.54_{\pm 0.37}$	4.81	$33.65_{\pm 2.84}$	$33.20_{\pm 3.16}$	$36.30_{\pm 1.12}$	2.65	$10.10_{\pm 2.66}$	$9.15_{\pm 1.68}$	$12.57_{\pm 2.94}$	2.47
FWD-R	$52.55_{\pm 4.06}$	$49.17_{\pm 3.11}$	$61.32_{\pm 0.90}$	8.77	$18.29_{\pm 0.39}$	$16.61_{\pm 2.65}$	$21.50_{\pm 0.38}$	3.21	$32.15_{\pm 3.40}$	$33.10_{\pm 2.03}$	$35.70_{\pm 1.19}$	2.60	$12.72_{\pm 3.28}$	$11.57_{\pm 2.91}$	$14.85_{\pm 1.75}$	2.12
URE-GA-U	$48.68_{\pm 1.11}$	$49.29_{\pm 1.67}$	$50.24_{\pm 1.11}$	0.95	$15.23_{\pm 2.35}$	$16.09_{\pm 1.23}$	$16.67_{\pm 1.35}$	0.58	$28.10_{\pm 5.24}$	$34.35_{\pm 2.39}$	$35.70_{\pm 1.97}$	1.35	$8.53_{\pm 1.55}$	$8.52_{\pm 1.38}$	$11.65_{\pm 1.90}$	3.12
URE-GA-R	$50.49_{\pm 1.21}$	$50.25_{\pm 1.57}$	$50.73_{\pm 1.83}$	0.25	$15.68_{\pm 1.35}$	$16.12_{\pm 0.95}$	$17.57_{\pm 0.61}$	1.45	$29.85_{\pm 4.73}$	$34.10_{\pm 1.90}$	$33.65_{\pm 1.40}$	-0.45	$7.15_{\pm 2.13}$	$7.12_{\pm 2.42}$	$9.78_{\pm 3.88}$	2.63
SCL-NL	$54.32_{\pm 6.71}$	$51.03_{\pm 3.12}$	$63.76_{\pm 0.09}$	9.44	$15.65_{\pm 3.06}$	$16.32_{\pm 3.11}$	$21.37_{\pm 1.18}$	5.05	$32.95_{\pm 3.13}$	$33.20_{\pm 3.69}$	$37.05_{\pm 1.40}$	3.85	$11.50_{\pm 3.76}$	$9.28_{\pm 2.55}$	$13.00_{\pm 2.80}$	1.50
SCL-EXP	$50.98_{\pm 6.83}$	$41.61_{\pm 3.52}$	$63.29_{\pm 1.02}$	12.30	$16.71_{\pm 2.72}$	$16.15_{\pm 2.55}$	$21.57_{\pm 1.13}$	4.86	$32.95_{\pm 2.91}$	$29.70_{\pm 2.83}$	$36.55_{\pm 1.28}$	3.60	$10.53_{\pm 2.02}$	$8.83_{\pm 3.19}$	$12.95_{\pm 3.38}$	2.43
L-W	$46.88_{\pm 9.44}$	$50.36_{\pm 0.47}$	$54.32_{\pm 0.41}$	3.95	$16.26_{\pm 1.93}$	$14.67_{\pm 1.59}$	$19.59_{\pm 0.99}$	3.33	$17.70_{\pm 9.90}$	$28.60_{\pm 5.15}$	$33.80_{\pm 2.66}$	5.20	$8.58_{\pm 1.25}$	$7.70_{\pm 0.35}$	$12.70_{\pm 2.35}$	4.12
L-UW	$52.47_{\pm 3.63}$	$51.15_{\pm 1.61}$	$57.52_{\pm 0.59}$	5.05	$16.10_{\pm 1.51}$	$15.58_{\pm 1.97}$	$20.71_{\pm 0.92}$	4.62	$22.10_{\pm 7.68}$	$25.60_{\pm 7.14}$	$35.10_{\pm 2.74}$	9.50	$10.60_{\pm 2.36}$	$8.28_{\pm 2.02}$	$12.12_{\pm 3.13}$	1.52
PC-sigmoid	$35.29_{\pm 1.67}$	$34.82_{\pm 1.24}$	$37.78_{\pm 0.80}$	2.49	$13.41_{\pm 0.95}$	$13.40_{\pm 0.72}$	$14.48_{\pm 0.47}$	1.07	$25.55_{\pm 5.99}$	$27.05_{\pm 5.66}$	$29.10_{\pm 0.98}$	2.05	$7.75_{\pm 1.73}$	$8.72_{\pm 0.26}$	$10.72_{\pm 1.38}$	2.00
ROB-MAE	$57.99_{\pm 1.72}$	$57.79_{\pm 2.03}$	$59.38_{\pm 0.63}$	1.39	$17.07_{\pm 2.02}$	$15.62_{\pm 1.79}$	$18.17_{\pm 1.31}$	1.11	$30.15_{\pm 4.22}$	$29.15_{\pm 2.90}$	$31.50_{\pm 1.81}$	1.35	$5.42_{\pm 0.27}$	$5.03_{\pm 0.54}$	$6.35_{\pm 0.86}$	0.92

Table 5: The testing accuracy of models evaluated with URE and SCEL.

	CLCIFAR10				CLCIFAR20				CLMIN10				CLMIN20			
	URE	SCEL	valid acc	gap (↓)	URE	SCEL	valid acc	gap (↓)	URE	SCEL	valid acc	gap (↓)	URE	SCEL	valid acc	gap (↓)
FWD-U	$33.13_{\pm 1.30}$	$31.86_{\pm 1.52}$	$34.83_{\pm 0.50}$	1.70	$6.70_{\pm 0.46}$	$7.10_{\pm 0.48}$	$8.03_{\pm 0.74}$	0.93	$20.75_{\pm 2.12}$	$20.20_{\pm 0.72}$	$23.85_{\pm 2.76}$	3.10	$4.97_{\pm 0.72}$	$4.55_{\pm 0.81}$	$6.33_{\pm 1.04}$	1.35
FWD-R	$33.70_{\pm 3.38}$	$35.64_{\pm 1.37}$	$38.13_{\pm 0.88}$	2.49	$17.35_{\pm 2.32}$	$18.40_{\pm 1.56}$	$20.27_{\pm 0.53}$	1.86	$22.15_{\pm 4.15}$	$29.15_{\pm 1.93}$	$30.15_{\pm 1.83}$	1.00	$8.60_{\pm 1.32}$	$9.90_{\pm 1.19}$	$10.60_{\pm 0.82}$	0.70
URE-GA-U	$30.45_{\pm 3.58}$	$33.21_{\pm 1.12}$	$34.72_{\pm 0.40}$	1.51	$7.03_{\pm 0.61}$	$8.71_{\pm 0.74}$	$10.49_{\pm 0.52}$	1.79	$17.05_{\pm 3.35}$	$21.30_{\pm 3.01}$	$22.90_{\pm 2.97}$	1.60	$4.27_{\pm 0.80}$	$5.03_{\pm 0.48}$	$5.75_{\pm 0.43}$	0.72
URE-GA-R	$27.39_{\pm 1.89}$	$28.32_{\pm 1.38}$	$30.23_{\pm 0.70}$	1.91	$3.58_{\pm 0.47}$	$5.42_{\pm 0.96}$	$6.17_{\pm 0.82}$	0.75	$8.90_{\pm 1.03}$	$10.30_{\pm 1.53}$	$13.25_{\pm 5.11}$	2.95	$5.15_{\pm 0.62}$	$5.57_{\pm 1.54}$	$6.50_{\pm 0.35}$	0.93
SCL-NL	$33.55_{\pm 0.79}$	$33.70_{\pm 1.33}$	$34.77_{\pm 0.60}$	1.07	$6.73_{\pm 0.51}$	$7.47_{\pm 0.56}$	$8.02_{\pm 0.36}$	0.55	$19.55_{\pm 1.37}$	$22.15_{\pm 1.76}$	$21.80_{\pm 1.85}$	-0.35	$4.83_{\pm 1.12}$	$5.20_{\pm 0.51}$	$6.17_{\pm 0.49}$	0.98
SCL-EXP	$31.30_{\pm 2.62}$	$33.47_{\pm 1.16}$	$35.18_{\pm 0.67}$	1.71	$6.83_{\pm 0.23}$	$7.03_{\pm 0.62}$	$7.70_{\pm 0.41}$	0.66	$18.35_{\pm 1.60}$	$20.65_{\pm 1.39}$	$24.80_{\pm 1.14}$	4.15	$5.05_{\pm 0.56}$	$4.45_{\pm 0.74}$	$5.58_{\pm 0.13}$	0.52
L-W	$27.49_{\pm 4.30}$	$30.32_{\pm 2.40}$	$32.99_{\pm 1.01}$	2.67	$5.90_{\pm 0.29}$	$7.18_{\pm 0.31}$	$7.71_{\pm 0.35}$	0.53	$19.30_{\pm 4.66}$	$18.95_{\pm 2.30}$	$23.80_{\pm 2.64}$	4.50	$5.97_{\pm 0.33}$	$5.55_{\pm 0.17}$	$6.40_{\pm 0.29}$	0.43
L-UW	$28.90_{\pm 2.01}$	$29.78_{\pm 2.69}$	$34.69_{\pm 0.32}$	4.91	$6.40_{\pm 0.42}$	$8.16_{\pm 0.30}$	$8.15_{\pm 0.30}$	-0.01	$18.25_{\pm 4.31}$	$19.80_{\pm 1.61}$	$22.40_{\pm 1.67}$	2.60	$5.82_{\pm 0.77}$	$6.48_{\pm 1.03}$	$6.35_{\pm 0.86}$	-0.13
PC-sigmoid	$24.83_{\pm 5.94}$	$31.48_{\pm 1.93}$	$32.15_{\pm 0.80}$	0.67	$7.98_{\pm 2.47}$	$10.59_{\pm 0.87}$	$12.11_{\pm 0.46}$	1.51	$12.55_{\pm 1.31}$	$17.85_{\pm 4.61}$	$23.15_{\pm 0.46}$	5.30	$6.40_{\pm 1.19}$	$5.33_{\pm 1.28}$	$6.90_{\pm 1.04}$	0.50
ROB-MAE	$18.80_{\pm 1.64}$	$18.75_{\pm 0.99}$	$20.23_{\pm 1.02}$	1.43	$4.70_{\pm 0.43}$	$4.87_{\pm 0.32}$	$5.40_{\pm 0.59}$	0.53	$11.80_{\pm 2.92}$	$14.35_{\pm 1.59}$	$14.15_{\pm 0.68}$	-0.20	$5.08_{\pm 0.44}$	$4.62_{\pm 0.66}$	$5.38_{\pm 0.33}$	0.30

Results and discussion Firstly, there appears no clear winner between URE and SCEL, both using only CLs for validation. Validating with the ordinary-label accuracy generally provides stronger performance than URE/SCEL, and the test performance gap between validating with ordinary labels and validating with complementary labels can be as big as nearly 5%. These findings suggest that using purely complementary labels for validation, whether through URE or SCEL, still suffers from a non-negligible performance drop compared to using ordinary validation. That is, the numbers reported in existing studies, which validates with ordinal labels, can be optimistic for practice. Whether this gap can be further reduced remains an open research problem and the community can pay more attention on that to make CLL more practical.

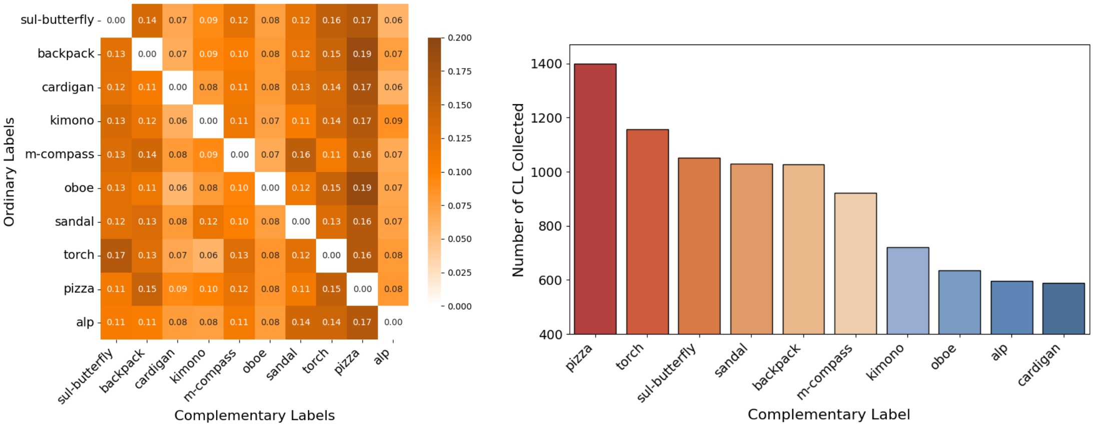
7 Alternative to Current Data Collection Protocol

In this work, we used a specific protocol to collect complementary labels, but we acknowledge there are several alternative protocols that could also be considered. Here we highlight possibilities while explaining the rationale behind our current design choices.

The first alternative would be to provide annotators with more than four classes. However, realistically, increasing the number beyond four would significantly increase the effort required from annotators. We chose four because it is a common standard for multiple-choice scenarios and was effectively employed in prior research (Wei et al., 2022) that used human annotations to collect noisy labels in real-world settings. Nevertheless, we admit exploring different numbers of labels beyond four presents an interesting area for future research in complementary dataset.

A second alternative, we considered was using a simple “yes/no” format. However, this approach posed two main issues. First, it would generate a dataset containing both true and complementary labels, which conflicts with our goal of creating a purely complementary dataset. Second, the resulting dataset would likely contain increased noise, complicating clear analysis and validation of the label transition assumptions.

Another possible protocol involves explicitly hiding the true label during annotation to minimize noise. However, this approach is less realistic since, in real-world scenarios, true labels are typically unknown (otherwise complementary annotation will not be necessary). Note that part of our current dataset, which



(a) Transition matrix (b) Label distribution

Figure 7: The transition matrix (**a**) and label distribution (**b**) of CLMicroImageNet10 when hidden the true label.

does not have choices that include the true label, can be taken to analyze the effect of such a protocol. We show in Figure 7) that such a protocol does not lead to much change of the transition matrix (except for the diagonal noise removal) and label distribution, when compared with our original choice of protocol.

We view our work as an initial exploration into collecting complementary labels, and we hope that this work will inspire the future work on other possible collection protocols of complementary label.

8 Conclusion

In this paper, we devised a protocol to collect complementary labels from human annotators. Utilizing this protocol, we curated four real-world datasets, CLCIFAR10, CLCIFAR20, CLMicroImageNet10, and CLMicroImageNet20 and made them publicly available to the research community. Through our meticulous analysis of these datasets, we confirmed the presence of noise and bias in the human-annotated complementary labels, challenging some of the underlying assumptions of existing CLL algorithms. Extensive benchmarking experiments revealed that noise is a critical factor that undermines the effectiveness of most existing CLL algorithms. Furthermore, the biased complementary labels can trigger overfitting, even for algorithms explicitly designed to leverage this bias information. In addition, our study on the validation objective for CLL suggests that validating with only complementary labels causes significant performance degrading. These findings emphasize the need for the community to dedicate more effort on those issues. The curated datasets pave the way for the community to create more practical and applicable CLL solutions.

9 Limitations

To ensure the compatibility with previous CLL algorithms, our work focuses on image datasets based on CIFAR10/100, and TinyImageNet. It is worth investigating the real-world CLL datasets on larger datasets, such as ImageNet, and other domains. On the other hand, the proposed protocol focuses on collecting real-world complementary labels for analyzing the common assumptions on CLL. That said, it is also crucial to understand efficient ways to collect complementary labels in practice, e.g., by asking annotators binary questions to collect ordinary and complementary labels simultaneously. We leave these directions as future works and hope that our work can open the way for the community to understand these questions.

Broader Impact Statement

It is known that weakly-supervised learning can be used for some privacy-preserving applications. While our CLImage datasets do not belong to those. We suggest practitioners exercise caution when studying such applications.

References

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels, 2020.

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

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019.

François Denis. Pac learning from positive statistical queries. In *Algorithmic Learning Theory: 9th International Conference, ALT'98 Otzenhausen, Germany, October 8–10, 1998 Proceedings 9*, pp. 112–126. Springer, 1998.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Yi Gao and Min-Ling Zhang. Discriminative complementary-label learning with weighted loss. In *International Conference on Machine Learning*, pp. 3587–3597. PMLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016.

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 Krishna Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models, 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.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

Wei-I Lin and Hsuan-Tien Lin. Reduction from complementary-label learning to probability estimates. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, May 2023.

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Consistent complementary-label learning via order-preserving losses. In *International Conference on Artificial Intelligence and Statistics*, pp. 8734–8748. PMLR, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: a loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3), 2015.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, Tomoya Sakai, and Gang Niu. *Machine learning from weak supervision: An empirical risk minimization approach.* MIT Press, 2022.

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

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations, 2022.

Yanwu Xu, Mingming Gong, Junxiang Chen, Tongliang Liu, Kun Zhang, and Kayhan Batmanghelich. Generative-discriminative complementary learning, 2019. URL `https://arxiv.org/abs/1904.01612`.

Nai-Xuan Ye, Tan-Ha Mai, Hsiu-Hsuan Wang, Wei-I Lin, and Hsuan-Tien Lin. libcll: an extendable python toolkit for complementary-label learning, 2024. URL `https://arxiv.org/abs/2411.12276`.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels, 2018.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.