Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 32(10): 5804–5818.
DOI: 10.3934/era.2024268
Received: 08 September 2024
Revised: 16 October 2024
Accepted: 16 October 2024
Published: 25 October 2024

Research article

Open-world barely-supervised learning via augmented pseudo labels

Zhongnian Li[1,2], **Yanyan Ding**[1], **Meng Wei**[1] **and Xinzheng Xu**[1,2,3,*]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

[2] Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China

[3] State Key Lab for Novel Software Technology, Nanjing University, Nanjing 220023, China

* **Correspondence:** Email: xxzheng@cumt.edu.cn.

Abstract: Open-world semi-supervised learning (OWSSL) has received significant attention since it addresses the issue of unlabeled data containing classes not present in the labeled data. Unfortunately, existing OWSSL methods still rely on a large amount of labeled data from seen classes, overlooking the reality that a substantial amount of labels is difficult to obtain in real scenarios. In this paper, we explored a new setting called open-world barely-supervised learning (OWBSL), where only a single label was provided for each seen class, greatly reducing labeling costs. To tackle the OWBSL task, we proposed a novel framework that leveraged augmented pseudo-labels generated for the unlabeled data. Specifically, we first generated initial pseudo-labels for the unlabeled data using visual-language models. Subsequently, to ensure that the pseudo-labels remained reliable while being updated during model training, we enhanced them using predictions from weak data augmentation. This way, we obtained the augmented pseudo-labels. Additionally, to fully exploit the information from unlabeled data, we incorporated consistency regularization based on strong and weak augmentations into our framework. Our experimental results on multiple benchmark datasets demonstrated the effectiveness of our method.

Keywords: open-world; barely-supervised learning; semi-supervised learning; CLIP; pseudo-label

1. Introduction

Semi-supervised learning (SSL) was introduced to address the problem of limited labeled data in machine learning. SSL uses a small number of labeled samples and tries to utilize a large number of unlabeled samples to improve the performance of the model. In many fields, such as image classification [1–3], semantic segmentation [4–6], and object detection [7–9], SSL methods have achieved outstanding results.

Regrettably, traditional SSL methods [10, 11] require labeled samples of all classes, which is almost impossible in real-world scenarios. Open-world SSL (OWSSL) was proposed, assuming that the unlabeled dataset not only contains the classes of labeled data, but also existing unseen categories that have never been seen in labeled datasets. The goal of OWSSL is to distinguish seen classes while also discovering unseen classes. Cao et al. [12] introduced an uncertainty adaptive margin mechanism to discover novel classes while avoiding bias toward seen classes. NACH [13] exploited pair-wise similarities between examples to discover unseen classes, while balancing the learning speed of seen and unseen classes via an adaptive threshold with distribution alignment.

However, existing OWSSL methods do not account for scenarios where the number of labeled samples in seen classes is extremely limited, which is quite likely to arise in practical applications. This inspires us to propose a new setting called open-world barely-supervised learning (OWBSL), where each seen class is represented by only one label. As shown in Figure 1, our OWBSL setting is illustrated in Figure 1(c). Unlike Figure 1(a), where both labeled and unlabeled data contain the same classes, and different from Figure 1(b), where seen classes have a large number of labeled samples, our proposed OWBSL setting features unlabeled samples containing classes not present in the labeled samples, while each seen class has only one labeled instance.

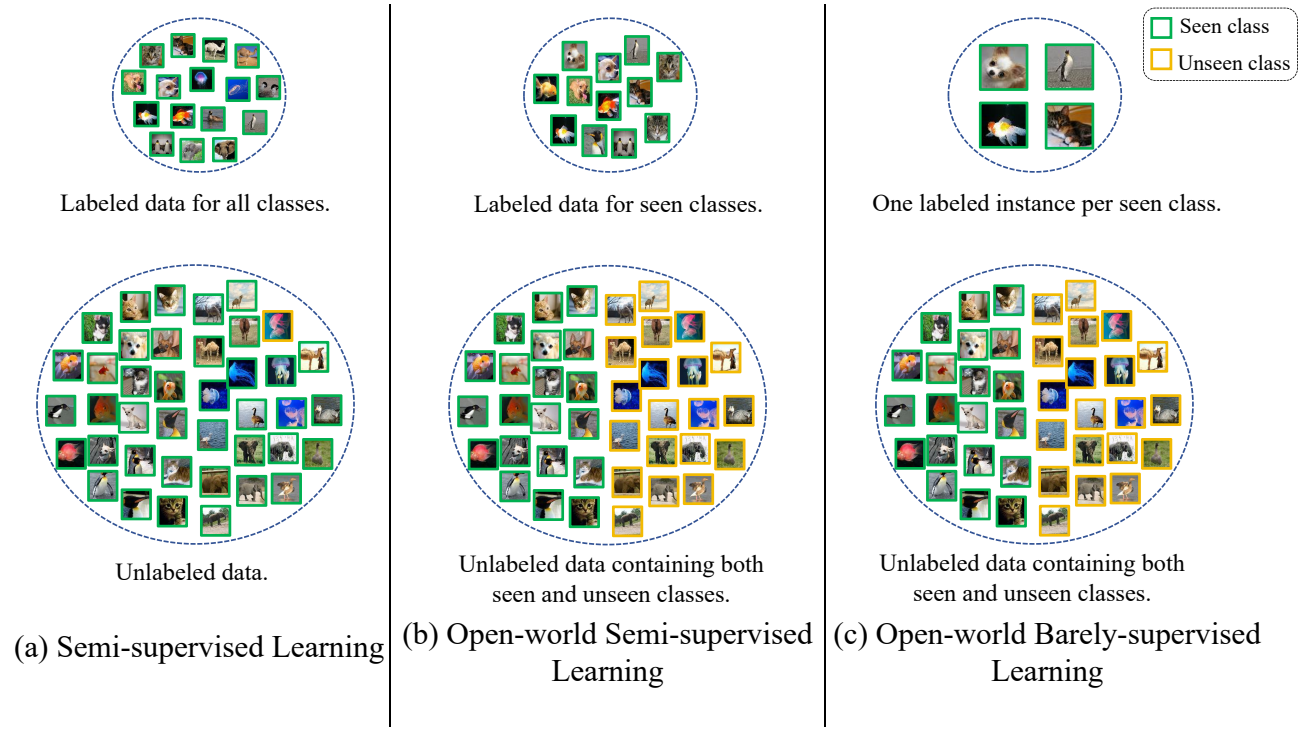

Figure 1. Comparison of OWBSL with OWSSL and SSL. (a) represents SSL with a small amount of labeled data and a large amount of unlabeled data, where labeled and unlabeled data follow the same distribution. (b) represents OWSSL, where the unlabeled data includes unseen classes not present in the labeled data. (c) represents our OWBSL setting, where each seen class has only one labeled instance.

Under the OWBSL setting, the standard OWSSL method often fails to work because it cannot obtain reliable labels. To tackle this challenge, we propose a new framework that incorporates contrastive

language-image pretraining (CLIP) [14] to provide relatively reliable pseudo-labels for unlabeled data as prior knowledge. Specifically, we first leverage CLIP's zero-shot inference capability to generate pseudo-labels for the unlabeled training data. Then, we enhance these pseudo-labels by incorporating data augmentation techniques. We use the predictions obtained from the model after applying weak augmentation to the images as a component of the pseudo-labels, combining them with CLIP's output to form the augmented pseudo-labels. Additionally, to improve the model's robustness and better distinguish both seen and unseen classes, we apply strong augmentations to the images and enforce consistency between the outputs of the strong and weak augmentations to guide the model's learning process.

We summarize our contributions as follows:

- We propose a new OWBSL setting where there is only one label for each seen class. In this setting, only a minimal number of labeled samples from the seen class are required to classify between seen and unseen classes.
- We propose a new framework that can provide reliable pseudo-labeling for unseen classes. We combine the zero-shot inference capability of vision-language models (VLMs) with data augmentation techniques to generate augmented pseudo-labels. Additionally, we incorporate data augmentation and consistency regularization techniques to help the model learn representations.
- Experimental results on four benchmark datasets demonstrate the effectiveness of our method.

2. Related work

2.1. Semi-supervised learning

Pseudo-labeling [15–19] and consistency regularization [20–23] are two trends of SSL. The core idea of pseudo-labeling is to use the model's prediction to label the unlabeled data, and then train the model using these pseudo-labels alongside the original labeled data. Consistency regularization is a technique that enforces consistency constraints on different perturbed versions of unlabeled data. FixMatch [24] proposed a simple framework which achieves excellent performance by unifying these two methods. It uses the model's high confidence predictions of weak data augmentation as pseudo-labels, forcing the model to learn consistency between the strong and weak augmentation of the data.

FlexMatch [25] used adaptive thresholds for each class, and, similarly, MarginMatch [10] and FreeMatch [11] have also designed different threshold selection strategies. PROTOCON [26] refined pseudo-labels through clustering by utilizing information from the sample's nearest neighbors. Chen et al. [27] argued that pseudo-labeling strategies based on threshold selection overlooked unlabeled data. To address this, they introduced entropy meaning loss and adaptive negative learning techniques to more effectively utilize unlabeled data. Robust SSL aims to improve the performance of models in the presence of noisy or inaccurate labels. Park et al. [28] effectively identified out-of-class data by assigning soft labels to out-of-class unlabeled data using self-supervised contrastive learning. Mo et al. [29] improved the generalization ability of the model by calibrating predictions through density modeling in the representation space.

Thomas et al. [30] proposed the method which is based on self-supervised clustering and selects pseudo-labels by utilizing historical predictions of samples and class-dependent thresholds. Gui et al. [31] posited that insufficient labels result in inadequate learning of discriminative information.

To address this, it proposes constructing super-classes and using the similarity between samples and super-classes to enhance the learning of discriminative information.

2.2. Open-world semi-supervised learning

Although the above SSL methods have achieved excellent results, they still have the limitation of requiring labeled data for all categories. To address this, a new scenario called OWSSL has been proposed, where the unlabeled data includes classes that are not present in the labeled data.

OpenCon [32] combined contrastive learning and proposed a prototype-based algorithm that could discover new classes through clustering without knowing the number of classes. Rizve et al. [33] utilized prior knowledge of class distributions to generate reliable class distribution aware pseudo-labels for unlabeled data. Taxonomic context priors discovering and aligning (TIDA) [34] exploited multi-granularity semantic concepts as prior knowledge to enhance representation learning and improve the quality of pseudo-labels.

Vaze et al. [35] used contrastive representation learning and clustering to directly provide class labels. Wen et al. [36] proposed a parameter classification method that benefits from entropy regularization, which achieved excellent performance. Zhao et al. [37] proposed an alternating learning framework that retrieves cluster assignments for unlabeled instances by identifying their nearest prototypes.

2.3. Vision-language models

VLMs integrate visual and textual information to enable comprehensive understanding and generation across modalities.

CLIP is trained on a massive amount of image-text pairs. Its core idea is to promote the correct pairing of images and texts through contrastive learning. CLIP demonstrates strong zero-shot learning capabilities. Context optimization (CoOp) [38] is based on CLIP and optimizes the context of the prompt word to make the model perform better in specific tasks. Conditional context optimization (CoCoOp) [39] argued that CoOp's generalization is insufficient and addresses this by generating input-conditional vectors for each image, enabling the model to generalize to unseen classes. CLIP Adapter [40] achieves better fine-tuning performance than CoOp by inserting the adapter module into CLIP's text encoder and image encoder. Tip-Adapter [41] leveraged the knowledge of few-shot samples by constructing a key-value cache model, enabling CLIP to adapt to few-shot classification tasks without requiring any parameter learning.

3. Methodology

3.1. Problem setting

Our OWBSL is defined as follows. The training dateset D contains two parts: labeled dataset $D_l = \{x_i, y_i\}_{i=1}^{n}$ and unlabeled dataset $D_u = \left\{u_j\right\}_{j=1}^{m}$. Typically, $m \gg n$. $x \in R^d$, where d is the feature dimension. $y \in \mathcal{Y}_l = \{c_i\}_{i=1}^{k_l}$, where k_l is the number of labeled classes. We use $\mathcal{Y}_{all} = \{c_i\}_{i=1}^{k_{all}}$ to represent all categories in the training set, $\mathcal{Y}_{seen}$ represents the seen classes, and $\mathcal{Y}_{unseen}$ represents the unseen classes. Specifically, $\mathcal{Y}_l = \mathcal{Y}_{seen} \subset \mathcal{Y}_{all}$, $\mathcal{Y}_{unseen} = \mathcal{Y}_{all} \setminus \mathcal{Y}_{seen}$ and $k_{all} > k_l$. Also, $\mathcal{Y}_{seen} \cap \mathcal{Y}_{unseen} = \varnothing$. In our setting, the number of labeled training data n is equal to the number of seen classes, that is, there is only one labeled example for each seen class.

3.2. Architecture

Overview As shown in Figure 2, our method mainly consists of two parts: supervised learning with labeled data and unsupervised learning with unlabeled data. The unsupervised part includes two main technologies: augmented pseudo-labeling and consistency regularization.

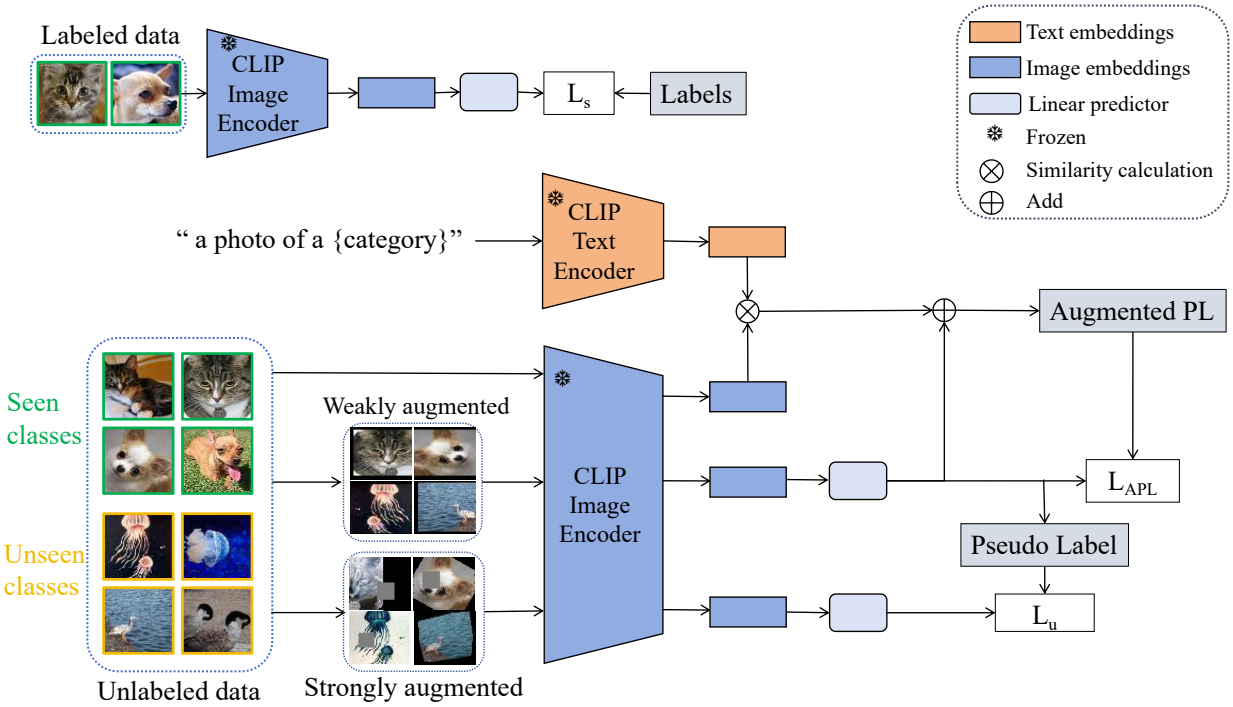

Figure 2. The overall architecture of our methods. Our approach mainly consists of two parts: supervised and unsupervised. The image and text encoders of CLIP are fixed, and "a photo of a [category]" is used as the text template. The overall loss of the model consists of supervised loss L_s and unsupervised losses L_{APL} and L_u.

Augmented pseudo-labels In this paper, we designed a pseudo-label enhancement method. Specifically, we introduce CLIP into OWBSL and use the prediction of CLIP as a component of pseudo-label. To fully leverage CLIP's zero-shot inference capability, we enhance the pseudo-labels using weak data augmentation techniques. These two components together form the augmented pseudo-labels.

Consistency regularization Consistency regularization leverages unlabeled data based on the assumption that the model should produce similar predictions when given different perturbations of the same image [24]. Therefore, to improve the model's stability, we continue to incorporate consistency regularization as a important component of our approach.

Supervised learning of labeled data For the labeled data D_l, we use the cross-entropy loss between the model's predicted probabilities and the true labels for training. Specifically, the cross-entropy loss function measures the difference between the model's predicted probability distribution and the actual label distribution, guiding the model to adjust its parameters to reduce prediction errors.

3.3. Augmented pseudo-labels

In this section, we explain how to generate augmented pseudo-labels for unlabeled data. Specifically, augmented pseudo-labels consist of two components: predictions from CLIP and predictions from weakly augmented data.

To start, we generate initial pseudo-labels using CLIP's zero-shot inference capabilities. CLIP is a multimodal model consisting of an image encoder $f^I(\cdot)$ and a text encoder $f^T(\cdot)$. For CLIP's text encoder, we use the template "a photo of a [category]" as the text description, where [category] is the name of the class. For one unlabeled sample u_i with a k-classification task, f_c^T is the text feature for the c-th class extracted from pretrained text encoder and $f^I(u_i)$ is the image feature extracted from pretrained image encoder. Therefore, the probability of u_i in the c-th class is:

$$p_i^c = \frac{exp(< f_c^T, f^I(u_i) > /\tau)}{\sum_{j=1}^{k} exp(< f_j^T, f^I(u_i) > /\tau)} \tag{3.1}$$

where $< \cdot >$ represents cosine similarity, and τ represents the temperature parameter. Meanwhile, we can obtain the prediction probability vector $P_i^{CLIP} = [p_i^1, p_i^2, ..., p_i^k]$. Then, we can obtain the pseudo-label $\hat{y}_i$ of u_i from CLIP:

$$\hat{y}_i = softmax(p_i^{CLIP}) \tag{3.2}$$

Despite using Eqs (3.1) and (3.2), we can already obtain relatively reliable pseudo-labels, as relying solely on the pseudo-labels generated by CLIP has its limitations. This is because $f^I(\cdot)$ and $f^T(\cdot)$ from CLIP are frozen, which means that the generated pseudo-labels are fixed and cannot be updated. To address this issue, we incorporate weakly augmented predictions to refine the pseudo-labels.

Recognizing the substantial benefits of data augmentation technology in model training, we have implemented it in our approach. Let u_i^w and u_i^s represent the weak augmentation and strong augmentation views of the same image u_i, and $g(\cdot)$ to denote the output of the fully connected layer. For an unlabeled image with random weak augmentation u_i^w, the prediction is:

$$p_i^{FC} = g(f^I(u_i^w)) \tag{3.3}$$

From Eqs (3.2) and (3.3), we obtain the final augmented pseudo-labels:

$$APL_i = argmax(\hat{y}_i + softmax(p_i^{FC})) \tag{3.4}$$

3.4. Training objective

For a batch containing n labeled instances and m unlabeled instances, our loss function is divided into three parts: supervised loss L_s, unsupervised loss L_{APL} based on our augmented pseudo-labels, and unsupervised loss L_u based on consistency regularization.

The supervised loss for labeled data is:

$$L_s = \frac{1}{n} \sum_{i=1}^{n} H(y_i, g(f^I(x_i))) \tag{3.5}$$

where $H(\cdot)$ refers to cross-entropy loss.

The loss between the weakly augmentation output and the augmented pseudo-label is:

$$L_{APL} = \frac{1}{m}\sum_{i=1}^{m} H(APL_i, g(f^I(u_i^w))) \tag{3.6}$$

The unsupervised loss introduced by the threshold-based consistency regularization is:

$$L_u = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(max(g(f^I(u_i^w))) > \tau) H(argmax(g(f^I(u_i^w))), g(f^I(u_i^s))) \tag{3.7}$$

where τ is used to filter the pseudo-labels. Following FixMatch, we set τ as 0.95.

In summary, the overall training objective is:

$$L = L_s + L_{APL} + L_u \tag{3.8}$$

4. Experiments

4.1. Experimental setup

Dataset We evaluated our proposed approach on both generic image classification (CIFAR-100 [42], Tiny-ImageNet [43], Caltech-101 [44]) and fine-grained datasets (Food-101 [45]). For each dataset, we select 50, 20 and 80% of the classes as seen classes, with the remaining classes considered unseen. That is, the divisions between seen and unseen classes are 5/5, 2/8, and 8/2, respectively. In seen classes, each class randomly selects one sample to form the labeled training set, while the remaining samples, along with all samples from unseen classes, form the unlabeled training set. In particular, for the Caltech-101 dataset, we follow CoOp's approach by excluding the 'BACKGROUND_Google' and 'Faces_easy' classes. We present the specific details of the datasets in Table 1, including the number of classes, the amount of training and testing data, and the text templates used for each dataset.

Table 1. The details of datasets used for evaluations.

Dataset	Classes	Train	Test	Prompt
CIFAR-100	100	50,000	10,000	"a photo of a [CLASS]."
Tiny-ImageNet	200	100,000	10,000	"a photo of a [CLASS]."
Caltech-101	100	5769	2473	"a photo of a [CLASS]."
Food-101	101	75,750	25,250	"a photo of a [CLASS], a type of food."

Implementation details For every experiment, we use the CLIP's image encoder as feature extractor, and ViT-L/14 as the backbone of CLIP's image encoder. We set the batch size to 256 for every dataset. Besides, we employ AdamW [46] as the optimizer with weight decay 0.05 and the learning rate is set as 1e-3. For all experiments, we set the number of epochs to 50. For data augmentation, we employ horizontal flipping and random cropping as weak augmentations, while using methods that involve more substantial modifications, such as RandAugment [47] and Cutout [48], as strong augmentations.

Comparison methods To demonstrate the effectiveness of our method from multiple perspectives, we compared it with various approaches, including the CLIP baseline [14], CLIP-based methods [14], and representative methods of OWSSL. The specific introductions are as follows:

- CLIP baseline [14]. The CLIP baseline methods include CLIP zero-shot classification and CLIP linear probe. For CLIP's zero-shot classification, we used only the test set for evaluation, while CLIP linear probes trained on all available training data.
- CLIP-based methods [14]. We conducted experiments across multiple dimensions of CLIP. First, we define three methods based on CLIP linear probe (CLIP LP): CLIP one-shot LP (CLIP OLP), CLIP partial-classes LP (CLIP PLP), and CLIP one-shot partial-classes LP (CLIP OPLP). Specifically, CLIP OLP uses a randomly selected sample from each class to train the linear probe, CLIP PLP uses all the training data from the seen classes to train the linear probe, and CLIP OPLP uses only one sample from each seen class to train the linear probe.
- Open-world semi-supervised methods. We compared our approach with the state-of-the-art OWSSL methods, ORCA [12], and NACH [13]. To ensure fairness in the experiments, we maintained consistent settings during the comparison. Specifically, we use CLIP's image encoder as the feature extractor for ORCA [12] and NACH [13].

Table 2. Accuracy comparison results across four benchmarks with 5/5 splits between seen and unseen classes.

	CIFAR-100			Food-101			Caltech-101			Tiny-ImageNet		
Methods	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen
	CLIP baseline											
CLIP ZERO-SHOT	76.37	77.64	75.10	92.53	92.90	92.17	91.75	87.37	96.53	73.73	75.40	72.06
CLIP LP	85.80	86.38	85.22	94.72	94.55	94.90	98.95	99.27	98.69	85.18	85.20	85.17
	CLIP-based methods											
CLIP OLP	39.86	42.56	37.16	60.32	58.55	62.05	79.90	79.81	80.00	42.61	43.52	41.70
CLIP PLP	45.93	91.06	-	47.66	96.28	-	51.67	98.99	-	44.98	89.96	-
CLIP OPLP	25.99	51.92	-	31.73	64.09	-	43.83	83.97	-	25.60	51.14	-
	Open-world semi-supervised methods											
ORCA	41.71	40.94	49.34	10.70	8.99	17.18	16.05	11.03	20.59	17.48	23.62	19.48
NACH	51.83	45.14	54.30	9.79	6.30	16.29	26.85	23.94	28.95	28.43	19.74	34.74
OURS	**80.56**	**81.66**	**79.50**	**93.86**	**93.78**	**93.93**	**95.59**	**94.27**	**97.07**	**78.89**	**81.46**	**76.32**

4.2. Comparison with state-of-the-arts

In this section, we present a comparison between our method and other state-of-the-art methods to demonstrate the effectiveness of our approach. The specific results of our experiments are shown in Tables 2–4. Table 2 shows the results with 50% of the classes used as seen classes. Table 3 presents the results with 20% of the classes as seen classes. Table 4 displays the results with 80% of the classes as seen classes.

Table 3. Accuracy comparison results across four benchmarks with 2/8 splits between seen and unseen classes.

	CIFAR-100			Food-101			Caltech-101			Tiny-ImageNet		
Methods	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen
	CLIP baseline											
CLIP ZERO-SHOT	76.37	80.80	75.26	92.53	92.52	92.54	91.75	94.59	90.78	73.73	75.70	73.24
CLIP LP	85.80	87.25	85.44	94.72	93.96	94.91	98.95	99.50	98.77	85.18	84.30	85.40
	CLIP-based methods											
CLIP OLP	39.86	36.35	40.74	60.32	60.66	60.24	79.90	85.41	78.13	42.61	45.60	42.28
CLIP PLP	19.05	95.25	-	19.22	97.08	-	25.35	99.68	-	18.78	93.90	-
CLIP OPLP	10.75	53.75	-	14.33	72.36	-	23.66	93.00	-	11.39	56.95	-
	Open-world semi-supervised methods											
ORCA	40.32	26.75	42.53	8.36	15.48	9.41	20.29	45.11	15.94	16.08	21.35	16.79
NACH	51.37	8.35	54.45	8.96	16.16	10.38	24.87	50.41	18.82	26.70	14.25	29.96
OURS	**80.17**	**84.10**	**79.19**	**93.77**	**93.68**	**93.80**	**95.15**	**98.25**	**94.09**	**78.53**	**82.55**	**77.53**

Table 4. Accuracy comparison results across four benchmarks with 8/2 splits between seen and unseen classes.

	CIFAR-100			Food-101			Caltech-101			Tiny-ImageNet		
Methods	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen
	CLIP baseline											
CLIP ZERO-SHOT	76.37	75.80	78.65	92.53	93.09	91.31	91.75	90.30	97.57	73.73	74.28	71.55
CLIP LP	85.80	85.29	87.85	94.72	94.86	94.23	98.95	99.08	98.29	85.18	85.19	85.15
	CLIP-based methods											
CLIP OLP	39.86	40.11	38.85	60.32	60.39	60.08	79.90	78.37	87.59	42.61	43.35	39.65
CLIP PLP	69.45	86.81	-	75.87	95.78	-	82.25	98.64	-	69.54	86.93	-
CLIP OPLP	34.21	42.76	-	49.17	62.08	-	71.05	85.21	-	35.87	44.84	-
	Open-world semi-supervised methods											
ORCA	46.97	48.20	61.45	10.30	5.24	21.58	19.61	13.14	32.36	18.42	16.45	22.70
NACH	54.98	52.10	72.20	10.47	5.59	25.52	32.35	21.39	40.63	30.86	23.65	43.55
OURS	**80.30**	**80.02**	**81.40**	**93.75**	**93.91**	**93.12**	**95.67**	**95.29**	**97.57**	**79.29**	**80.28**	**75.40**

The comparison results show that our method achieves the best performance across all four datasets. To begin, it can be observed that our method outperforms CLIP's zero-shot classification across all datasets and shows only a small gap compared to fully supervised CLIP linear probes. Additionally, for CLIP-based methods such as CLIP OLP, CLIP PLP, and CLIP OPLP, our method significantly outperforms them. This also demonstrates that when the amount of data is very limited or the training data classes are incomplete, relying solely on the encoder and linear layers is insufficient. Finally, the experimental results clearly show that the drastic reduction in the number of labeled samples leads to a catastrophic decline in accuracy for ORCA [12] and NACH [13]. For example, as shown in Table 2, our method achieves the highest improvement of over 80% for all classes and seen classes, and the highest improvement of over 70% for unseen classes (Food-101). The significant gap between their results and ours further demonstrates the critical importance of semantic information.

4.3. Comparison of different split ratios

As previously mentioned, our experiments were conducted with different partition ratios between seen and unseen classes. Experimental results show that, for our method, as the proportion of seen classes increases, the accuracy across all classes also rises or remains relatively stable. Overall, despite varying partition ratios, our method maintains stability.

For the CLIP baseline method and CLIP OLP, since they use all categories during training, the differences in performance between seen and unseen classes are determined by the characteristics of the dataset itself. The accuracy of CLIP OLP and CLIP OPLP across all classes increases with the proportion of seen classes.

It can also be observed that OWSSL methods are significantly impacted by the proportion of seen classes. For instance, comparing Tables 3 and 4, we observe that NACH exhibits a discrepancy of over 40% in the accuracy of seen classes on CIFAR-100 when the seen/unseen class spilts are 2/8 and 8/2.

4.4. AUROC results

Area under the receiver operating characteristic curve (AUROC) is a commonly used open-set evaluation metric [49] to assess a model's performance across different thresholds. In our experimental setup, we treat seen classes as positive samples and unseen classes as negative samples. The results of our method and the comparison methods on the CIFAR-100 and Caltech-101 dataset are shown in Table 5 where we set the split of seen/unseen classes to 5/5. From the results, it can be observed that our method achieves excellent results, only slightly lower than the CLIP linear probe trained with all training samples. This indicates that our method effectively separates the seen and unseen classes.

Table 5. AUROC results on CIFAR-100 and Caltech-101 with 5/5 splits between seen and unseen classes.

Methods	CIFAR-100	Caltech-101
CLIP ZERO-SHOT	88.49	97.73
CLIP LP	98.48	99.98
CLIP OLP	78.65	96.41
CLIP PLP	87.01	97.82
CLIP OPLP	64.23	71.02
ORCA	74.68	51.26
NACH	82.32	60.07
OURS	97.22	99.87

4.5. Ablation study

To demonstrate the effectiveness of our method, we performed ablation experiments on the CIFAR-100 dataset. As can be seen from the Table 6, our method achieved the highest accuracy. It is evident that the results when only using CLIP prediction are the lowest, as the pseudo-labels are fixed and the model cannot learn from iterations. Integrating outputs from the weak augmentation branch and consistency regularization both enhance the model's accuracy, with the weak augmentation providing greater benefits. When all components are combined, as in our proposed method, the best results are achieved. Compared to the lowest results, our proposed method improves accuracy by 2.49% for all

classes, 2.6% for seen classes, and 2.86% for unseen classes.

Table 6. Ablation study results on CIFAR-100. Here, ✓ indicates that the corresponding component is included in the current experiment.

CLIP Prediction	Weakly Augmented Prediction	Consistency Regularization	All	Seen	Unseen
✓			78.07	79.06	77.06
✓	✓		79.43	80.38	78.48
✓		✓	78.81	80.98	76.64
✓	✓	✓	80.56	81.66	79.50

Additionally, we conducted experiments on the backbone of the CLIP image encoder to further validate the effectiveness of our method. As shown in Table 7, our experiments were performed on the Caltech-101 dataset. Even when using ViT-B/32 and ViT-B/16 as the backbones, our method still achieved results superior to zero-shot performance.

Table 7. Comparison experiments of backbone networks on the Caltech-101 dataset.

Backbone	CLIP ZERO-SHOT	All	Seen	Unseen
ViT-B/32	88.52	90.29	90.84	89.71
ViT-B/16	88.96	92.52	94.21	90.71

5. Conclusions

In this paper, we propose a new setting called OWBSL, where there is only one labeled data per seen class. At the same time, we propose a novel framework to solve this problem. We introduce CLIP in OWBSL to help generate augmented pseudo-labels. In order to enhance the discriminability of the model, we also utilize consistency regularization. Experimental results demonstrate that our method can effectively reduce the dependence on labeled data.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61976217 and 62306320), the Open Project Program of State Key Lab for Novel Software Technology (No. KFKT2024B32), and the Natural Science Foundation of Jiangsu Province (No. BK20231063).

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C. A. Raffel, Mixmatch: A holistic approach to semi-supervised learning, *Adv. Neural Inf. Process. Syst.*, **32** (2019).
2. D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, et al., Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring, preprint, arXiv:1911.09785. https://doi.org/10.48550/arXiv.1911.09785
3. Z. Peng, S. Tian, L. Yu, D. Zhang, W. Wu, S. Zhou, Semi-supervised medical image classification with adaptive threshold pseudo-labeling and unreliable sample contrastive loss, *Biomed. Signal Process. Control*, **79** (2023), 104142. https://doi.org/10.1016/j.bspc.2022.104142
4. Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, et al., Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 4238–4247. https://doi.org/10.1109/CVPR52688.2022.00421
5. H. Xu, L. Liu, Q. Bian, Z. Yang, Semi-supervised semantic segmentation with prototype-based consistency regularization, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 26007–26020.
6. H. Mai, R. Sun, T. Zhang, F. Wu, RankMatch: Exploring the better consistency regularization for semi-supervised semantic segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), 3391–3401. https://doi.org/10.1109/CVPR52733.2024.00326
7. H. Wang, Z. Zhang, J. Gao, W. Hu, A-teacher: Asymmetric network for 3D semi-supervised object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), 14978–14987. https://doi.org/10.1109/CVPR52733.2024.01419
8. M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, et al., End-to-end semi-supervised object detection with soft teacher, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 3060–3069. https://doi.org/10.1109/ICCV48922.2021.00305
9. J. Zhang, X. Lin, W. Zhang, K. Wang, X. Tan, J. Han, et al., Semi-detr: Semi-supervised object detection with detection transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 23809–23818. https://doi.org/10.1109/CVPR52729.2023.02280
10. T. Sosea, C. Caragea, MarginMatch: Improving semi-supervised learning with pseudo-margins, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 15773–15782. https://doi.org/10.1109/CVPR52729.2023.01514
11. Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, et al., FreeMatch: Self-adaptive thresholding for semi-supervised learning, in *The Eleventh International Conference on Learning Representations*, 2023.
12. K. Cao, M. Brbic, J. Leskovec, Open-world semi-supervised learning, preprint, arXiv:2102.03526. https://doi.org/10.48550/arXiv.2102.03526
13. L. Guo, Y. Zhang, Z. Wu, J. Shao, Y. Li, Robust semi-supervised learning when not all classes have labels, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 3305–3317.
14. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, in *International Conference on Machine Learning*, **139** (2021), 8748–8763.

15. D. H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in *Workshop on Challenges in Representation Learning*, ICML, **3** (2013), 896.

16. P. Cascante-Bonilla, F. Tan, Y. Qi, V. Ordonez, Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 6912–6920. https://doi.org/10.1609/aaai.v35i8.16852

17. J. Hu, C. Chen, L. Cao, S. Zhang, A. Shu, J. Jiang, et al., Pseudo-label alignment for semi-supervised instance segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), 16337–16347. https://doi.org/10.1109/ICCV51070.2023.01497

18. J. Li, C. Xiong, S. C. Hoi, Comatch: Semi-supervised learning with contrastive graph regularization, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 9475–9484. https://doi.org/10.1109/ICCV48922.2021.00934

19. E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, K. McGuinness, Pseudo-labeling and confirmation bias in deep semi-supervised learning, in *Proceedings of the 2020 International Joint Conference on Neural Networks*, (2020), 1–8. https://doi.org/10.1109/ijcnn48605.2020.9207304

20. S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, preprint, arXiv:1610.02242. https://doi.org/10.48550/arXiv.1610.02242

21. A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, *Adv. Neural Inf. Process. Syst.*, **30** (2017).

22. Q. Xie, Z. Dai, E. Hovy, T. Luong, Q. Le, Unsupervised Data Augmentation for Consistency Training, *Adv. Neural Inf. Process. Syst.*, **33** (2020), 6256–6268.

23. Y. Fan, A. Kukleva, D. Dai, B. Schiele Revisiting consistency regularization for semi-supervised learning, *Int. J. Comput. Vision*, **131** (2023), 626–643. https://doi.org/10.1007/s11263-022-01723-4

24. K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, et al., Fixmatch: Simplifying semi-supervised learning with consistency and confidence, *Adv. Neural Inf. Process. Syst.*, **33** (2020), 596–608.

25. B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, et al., Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 18408–18419.

26. I. Nassar, M. Hayat, E. Abbasnejad, H. Rezatofighi, G. Haffari, Protocon: Pseudo-label refinement via online clustering and prototypical consistency for efficient semi-supervised learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 11641–11650. https://doi.org/10.1109/CVPR52729.2023.01120

27. Y. Chen, X. Tan, B. Zhao, Z. Chen, R. Song, J. Liang, et al., Boosting semi-supervised learning by exploiting all unlabeled data, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 7548–7557. https://doi.org/10.1109/CVPR52729.2023.00729

28. J. Park, S. Yun, J. Jeong, J. Shin, Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data, in *European Conference on Computer Vision*, (2022), 134–149. https://doi.org/10.1007/978-3-031-25063-7_9

29. S. Mo, J. Su, C. Ma, M. Assran, I. Misra, L. Yu, et al., Ropaws: Robust semi-supervised representation learning from uncurated data, preprint, arXiv:2302.14483. https://doi.org/10.48550/arXiv.2302.14483

30. T. Lucas, P. Weinzaepfel, G. Rogez, Barely-supervised learning: Semi-supervised learning with very few labeled images, in *Thirty-Sixth AAAI Conference on Artificial Intelligence*, (2022), 1881–1889. https://doi.org/10.1609/aaai.v36i2.20082

31. G. Gui, Z. Zhao, L. Qi, L. Zhou, L. Wang, Y. Shi, Improving barely supervised learning by discriminating unlabeled samples with super-class, *Adv. Neural Inf. Process. Syst.*, **35** (2022), 19849–19860.

32. Y. Sun, Y. Li, Opencon: Open-world contrastive learning, preprint, arXiv:2208.02764. https://doi.org/10.48550/arXiv.2208.02764

33. M. N. Rizve, N. Kardan, M. Shah, Towards realistic semi-supervised learning, in *European Conference on Computer Vision*, (2022), 437–455. https://doi.org/10.1007/978-3-031-19821-2_25

34. Y. Wang, Z. Zhong, P. Qiao, X. Cheng, X. Zheng, C. Liu, et al., Discover and align taxonomic context priors for open-world semi-supervised learning, *Adv. Neural Inf. Process. Syst.*, **36** (2024).

35. S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Generalized category discovery, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 7492–7501. https://doi.org/10.1109/CVPR52688.2022.00734

36. X. Wen, B. Zhao, X. Qi, Parametric classification for generalized category discovery: A baseline study, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), 16590–16600. https://doi.org/10.1109/ICCV51070.2023.01521

37. B. Zhao, X. Wen, K. Han, Learning semi-supervised gaussian mixture models for generalized category discovery, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), 16623–16633. https://doi.org/10.1109/ICCV51070.2023.01524

38. K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, *Int. J. Comput. Vision*, **130** (2022), 2337–2348. https://doi.org/10.1007/s11263-022-01653-1

39. K. Zhou, J. Yang, C. C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 16816–16825. https://doi.org/10.1109/CVPR52688.2022.01631

40. P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, et al., Clip-adapter: Better vision-language models with feature adapters, *Int. J. Comput. Vision*, **132** (2024), 581–595. https://doi.org/10.1007/s11263-023-01891-x

41. R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, et al., Tip-adapter: Training-free clip-adapter for better vision-language modeling, preprint, arXiv:2111.03930. https://doi.org/10.48550/arXiv.2111.03930

42. A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, Master's thesis, University of Tront, 2009.

43. Y. Le, X. Yang, Tiny ImageNet Visual Recognition Challenge, *CS 231N*, **7** (2015), 3.

44. F. Li, F. Rob, P. Pietro, Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories, in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, (2004), 178–178. https://doi.org/10.1016/j.cviu.2005.09.012

45. L. Bossard, M. Guillaumin, L. V. Gool, Food-101-mining discriminative components with random forests, in *ECCV 2014*, (2014), 446–461. https://doi.org/10.1007/978-3-319-10599-4_29

46. I. Loshchilov, F. Hutter, Decoupled weight decay regularization, preprint, arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101

47. E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2020), 702–703. https://doi.org/10.1109/CVPRW50498.2020.00359

48. T. DeVries, Improved regularization of convolutional neural networks with cutout, preprint, arXiv:1708.04552. https://doi.org/10.48550/arXiv.1708.04552

49. H. Wang, G. Pang, P. Wang, L. Zhang, W. Wei, Y.Zhang, Glocal energy-based learning for few-shot open-set recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 7507–7516. https://doi.org/10.1109/CVPR52729.2023.00725

AIMS Press