Open-World Semi-Supervised Learning with Class Semantic Correlations

Yuxin Fan , **Junbiao Cui** , **Jiye Liang** and **Jianqing Liang** *

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China

fanyuxin@sxu.edu.cn, cjb@sxu.edu.cn, ljy@sxu.edu.cn, liangjq@sxu.edu.cn

Abstract

Open-world semi-supervised learning (OWSSL) aims to recognize both known and unknown classes, but the labeled samples only cover the known classes. Existing OWSSL methods primarily represent classes as symbolic variables, which ignore the rich internal semantic information associated with the classes and thus hampers their ability to recognize unknown classes. Recent studies incorporate textual descriptions of classes to facilitate training, but these methods overlook the class semantic correlations, which constrains their effectiveness in recognizing unknown classes. To address these issues, we propose a novel OWSSL method. Our method fine-tunes only the image encoder during training while keeping the text encoder frozen, thereby preserving the rich semantic correlations learned during the pre-training phase. Furthermore, we employ a semantic margin to extract class semantic correlations from textual descriptions, which are then utilized in enhancing image representation discriminability. Experimental results across multiple datasets demonstrate that our method significantly outperforms representative OWSSL methods in the recognition of both known and unknown classes.

1 Introduction

In the field of machine learning [LeCun *et al.*, 2015; Tu *et al.*, 2024] , traditional semi-supervised learning (SSL) frameworks [Sohn *et al.*, 2020] have provided an effective solution for scenarios with limited labeled samples. However, these frameworks typically operate under a strict assumption that the labeled samples encompass all classes within the application context. In real-world scenarios, especially in open environments where samples continuously arrive and unknown classes emerge, this assumption often does not hold [Li *et al.*, 2021] . This challenge gives rise to the concept of open-world semi-supervised learning (OWSSL) [Cao *et al.*, 2022] , which transcends the constraints of traditional SSL frameworks and aligns more closely with the complexities of the real-world. It requires models to be capable of effectively recognizing both known and unknown classes in unlabeled samples, even when the labeled samples only cover the known classes. Figure 1 illustrates the scenario encountered by OWSSL.

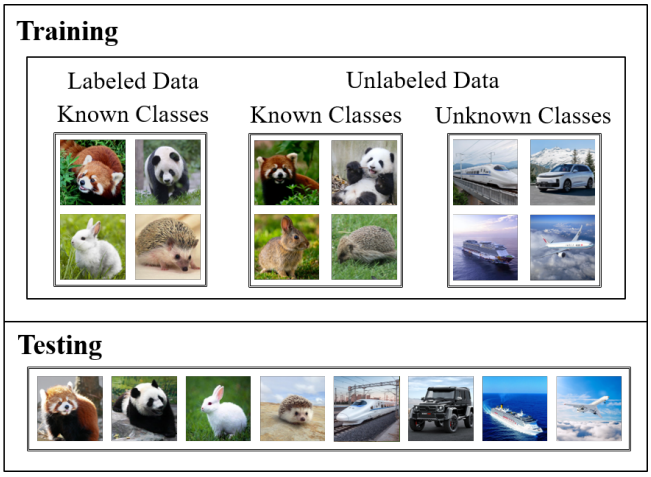

Figure 1: The environment faced by OWSSL. The model needs to be able to effectively recognize both known and unknown classes after training.

OWSSL is currently receiving widespread attention with a range of methods being proposed. For instance, ORCA [Cao *et al.*, 2022] introduces an adaptive margin mechanism to mitigate the model bias towards known classes, OpenLDN [Rizve *et al.*, 2022] proposes a bi-level optimization rule to generate relatively reliable pseudo-labels for unknown classes, and PromptCAL [Zhang *et al.*, 2023] proposes a two-stage framework to tackle the class collision issue caused by false negatives. The methods described above make significant progress, but they generally face a common issue: they primarily represent classes as symbolic variables. This representation can only indicate a sample's membership in a certain class and thus ignores the rich internal semantic information associated with the classes. This simplification limits the model to learning only the simple correspondence between samples and class symbols, resulting in weakly discriminative image representations, thereby impairing its ability to recognize unknown classes.

To address these issues, recent studies assume knowledge

*Corresponding author

of all class names and provide corresponding textual descriptions to fine-tune both image and text encoders, aiming to better understand the correspondence among visual features of samples and textual descriptions of classes. CLIP-GCD [Ouldnoughi *et al.*, 2023] first introduces textual descriptions and uses image-text contrastive loss to fine-tune the pre-trained CLIP model [Radford *et al.*, 2021]; TextGCD [Zheng *et al.*, 2024] then introduces a teacher model to mitigate the image-text matching noise in CLIP-GCD and has become the state-of-the-art in OWSSL. However, these methods do not consider the class semantic correlations when aligning the visual features of samples with their corresponding textual descriptions, resulting in insufficiently discriminative image representations. Moreover, fine-tuning the text encoder may disrupt the rich semantic correlations learned during pre-training phase, potentially leading to overfitting to the provided textual descriptions [Shu *et al.*, 2023]. Both of these issues consequently affect the model's ability in recognizing unknown classes.

In light of the issues of existing methods, we propose a novel OWSSL method. This method fine-tunes only the image encoder during training while freezing the text encoder, thereby preserving the rich semantic correlations learned during the pre-training phase. Furthermore, our method employs a semantic margin to extract class semantic correlations from textual descriptions, which are then used in enhancing image representation discriminability. Experimental results demonstrate that our method achieves a significant performance improvement compared to representative OWSSL methods. The main contributions of this paper are as follows:

- We identify the key issue in existing OWSSL methods that utilize textual descriptions: they do not consider the class semantic correlations when aligning the visual features of samples with their corresponding textual descriptions, resulting in insufficiently discriminative image representations, thereby affecting their ability to recognize unknown classes.
- We propose a novel OWSSL method named CSC-OWSSL, which freezes the text encoder, thereby preserving the rich semantic correlations learned during the pre-training phase and employs a semantic margin to extract class semantic correlations from textual descriptions, which are then used in enhancing image representation discriminability.
- We conduct comprehensive experiments on multiple fine-grained datasets. The experimental results demonstrate that our method achieves a significant performance improvement compared to representative OWSSL methods, fully illustrating the effectiveness of our method.

2 Related Work

2.1 Open-World Machine Learning

Most machine learning methods operate under the closed-world assumption, which frequently proves inadequate in real-world scenarios. To address this, methods such as open-set recognition (OSR) [Scheirer *et al.*, 2013] , robust semi-supervised learning (Robust SSL) [Oliver *et al.*, 2018] , and novel class discovery (NCD) [Han *et al.*, 2019] have been developed for open-world adaptation. OSR requires models to recognize unseen classes during testing without compromising known class accuracy. Robust SSL assumes that unlabeled samples may contain classes not represented in the labeled samples, and the goal is to minimize any negative impact of unknown classes on the performance of known class classification. NCD assumes that the unlabeled samples consist solely of unknown classes, requiring the model to recognize these novel classes. Although these methods make progress, they do not fully capture the complexity of open-world scenarios. To effectively address open-world machine learning challenges, OWSSL is proposed to better align with real-world complexities.

2.2 Open-World Semi-Supervised Learning

[Cao *et al.*, 2022] first proposes the concept of OWSSL, which requires the model to recognize both known and unknown classes simultaneously, thereby significantly enhancing the flexibility and generalization capabilities of SSL. [Vaze *et al.*, 2022] introduces the concept of Generalized Category Discovery (GCD), which is similar to the concept of OWSSL and is often discussed alongside it. Existing OWSSL methods can primarily be categorized into two types: one type represents classes as symbolic variables, with representative methods such as SimGCD [Wen *et al.*, 2023] and GPC [Zhao *et al.*, 2023]; the other type incorporates textual descriptions of classes, with representative methods such as CLIP-GCD [Ouldnoughi *et al.*, 2023] and TextGCD [Zheng *et al.*, 2024]. To broaden the application scope of these methods, some researchers extend them to various fields such as point cloud segmentation [Riz *et al.*, 2023] and intent classification [Shi *et al.*, 2024; An *et al.*, 2024] . Nevertheless, most research still focuses on image classification tasks, as they are crucial scenarios for testing the model's generalization ability and its capability to handle unknown classes. Therefore, in this paper, we choose to apply and validate the effectiveness of our proposed methods on image classification tasks.

2.3 Vision-Language Models

In recent years, Visual-Language Models (VLMs) have made significant advancements. These models are designed to process and understand visual and textual information. The main objective in this field is to train VLMs with a vast number of image-text pairs, allowing them to capture the correlation between visual content and textual descriptions. These models exhibit exceptional performance across various visual-language tasks [Zang *et al.*, 2024]. Specifically, CLIP [Radford *et al.*, 2021] attains impressive capabilities in understanding cross-modal concepts and correlations by aligning image and text features in a shared latent space, after being trained on extensive datasets. This robust generalization ability, along with the model's capability to handle multi-modal information, makes CLIP suitable for OWSSL scenarios, as it can adapt to and manage a broad spectrum of visual and textual inputs, which is essential for OWSSL methods.

3 Proposed Method

In this section, we first describe the OWSSL setting. Subsequently, we introduce each part of our method. The overall framework of our method is depicted in Figure 2 .

3.1 Problem Setting

Given a labeled dataset D_l containing n samples, $D_l = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, and an unlabeled dataset D_u containing m samples, $D_u = \{x_{n+1}, x_{n+2}, \cdots, x_{n+m}\}$, it is usually assumed that $m \gg n$. Here, $x \in \mathbb{R}^D$, $y \in C_l$, where D represents the feature dimensions of the samples. C_l denotes the set of classes in D_l and C_u denotes the set of classes in D_u, $C = C_l \cup C_u$. In OWSSL, $C_{known} = C_l$ represents the set of known classes and $C_{unknown} = C_u \backslash C_l$ represents the set of unknown classes. Our goal is to learn a classification model f to classify samples into known and unknown classes. The model is trained on D_l and D_u, and then evaluated on D_u. Knowing the number of unknown classes is a fundamental assumption in most OWSSL methods.

In this paper, we utilize deep neural networks to build our classification model $f(x; \theta_f)$, which consists of an image encoder $I(x; \theta_I) : \mathbb{R}^D \rightarrow \mathbb{R}^d$, a classifier $h(I(x; \theta_I); \theta_h) : \mathbb{R}^d \rightarrow \mathbb{R}^{|\mathbb{C}|}$, and a text encoder $T(t; \theta_T) : \mathbb{R}^D \rightarrow \mathbb{R}^d$. Here, D denotes the feature dimension of x, d represents the feature dimension following the processing by the image encoder, and t represents the textual descriptions. The dataset is processed in batches on the model, where B indexes all samples in a given batch, B_l indexes all labeled samples within that batch.

3.2 Baseline

We first establish a baseline model, which consist of the cross-entropy loss for labeled samples and the self-distillation loss [Assran *et al.*, 2022] for all samples. These loss functions are widely used in OWSSL methods [Cao *et al.*, 2022; Wen *et al.*, 2023]. The cross-entropy loss on labeled samples is as follows:

$$L_{sup} = \frac{1}{|B_l|} \sum_{i \in B_l} H\left(y_i, \sigma\big(p\big(x_i\big), \tau_s\big)\right), \tag{1}$$

where H denotes the cross-entropy function, $p\big(x_i\big) = h\left(I(x; \theta_I); \theta_h\right)$, $\sigma(\cdot)$ is the softmax function, and τ_s is a temperature value. The self-distillation loss on all samples is as follows:

$$L_{unsup} = \frac{1}{|B|} \sum_{i \in B} H\Big(\sigma\Big(p\big(x_i'\big), \tau_t\Big), \sigma\Big(p\big(x_i\big), \tau_s\Big)\Big) - R\big(\overline{p}\big), \tag{2}$$

where τ_t is a temperature value. $R\big(\overline{p}\big)$ is a mean-entropy maximization regularization term designed to prevent trivial solutions during initial training processing [Cao *et al.*, 2022], and $\overline{p} = \frac{1}{2|B|} \sum_{i \in B} \sigma\big(p\big(x_i\big), \tau_s\big) + \sigma\big(p\big(x_i'\big), \tau_s\big)$ is the mean softmax probability of a batch. Finally, the baseline loss is as follows:

$$L_{base} = \alpha L_{sup} + (1 - \alpha) L_{unsup}, \tag{3}$$

where α is a hyper-parameter, which is set to 0.35 by default.

However, L_{base} represents classes as symbolic variables, which can only indicate a sample's membership in a certain class and thus ignore the rich internal semantic information associated with the classes. This simplification limits the model to learning the simple correspondence between samples and class symbols, which hampers its ability to recognize unknown classes. Therefore, we need to incorporate corresponding textual descriptions of classes to facilitate training.

3.3 Semantic Margin Contrastive Loss

Recent OWSSL methods [Ouldnoughi *et al.*, 2023; Zheng *et al.*, 2024] assume knowledge of all class names and provide corresponding textual descriptions to fine-tune both image and text encoders, aiming to better understand the correspondence among visual features of samples and textual descriptions of classes. These methods fine-tune both image and text encoders using the image-text contrastive loss, denoted as L_{itc}, which is defined as follows:

$$L_{itc} = -\frac{1}{|B|} \sum_{i \in B} \log \frac{\exp\left(\frac{S\left(I(x_i; \theta_I), T\left(t_{y_i}; \theta_T\right)\right)}{\tau_k}\right)}{\sum_{k=1}^{|C|} \exp\left(\frac{S(I(x_i; \theta_I), T(t_k; \theta_T))}{\tau_k}\right)}, \tag{4}$$

where S is a similarity metric, specifically the cosine similarity, τ_k denotes the temperature value, and t represents the textual descriptions generated by a large language model (LLM). t_{y_i} represents the textual description associated with the sample x_i. It is retrieved by an auxiliary VLM from a knowledge base that encompasses textual descriptions for all classes, and it is the description that, after being processed by the text encoder, exhibits the highest similarity to x_i as processed by the image encoder. For more detailed information on the implementation, please refer to TextGCD [Zheng *et al.*, 2024]. Moreover, the class label y_i corresponding to t_{y_i} is also used as supervisory information within L_{base} to guide the training process of the model. However, these methods only align the visual features of samples with their corresponding textual descriptions, yet treat all non-corresponding textual descriptions equally, neglecting the class semantic correlations. Furthermore, fine-tuning the text encoder may disrupt the rich semantic correlations learned during pre-training phase, potentially leading to overfitting to the provided textual descriptions. Both of these issues consequently affect the model's generalization ability in recognizing unknown classes. To address these issues, we take the following steps:

First, we fine-tune only the image encoder during training while keeping the text encoder frozen. This approach preserves the rich semantic correlations that were learned during the pre-training phase. Furthermore, we employ a semantic margin to extract class semantic correlations from these descriptions [Shu *et al.*, 2023] and incorporate them into the image-text contrastive loss to fine-tune the image encoder, thereby enhancing image representation discriminability. The semantic margin is as follows:

$$M_{i,j} = 1 - S\left(T\left(t_i; \theta_T\right), T\left(t_j; \theta_T\right)\right), \tag{5}$$

where $i, j \in \{1, 2, \cdots |C|\}$.

Based on the semantic margin, the semantic margin con-

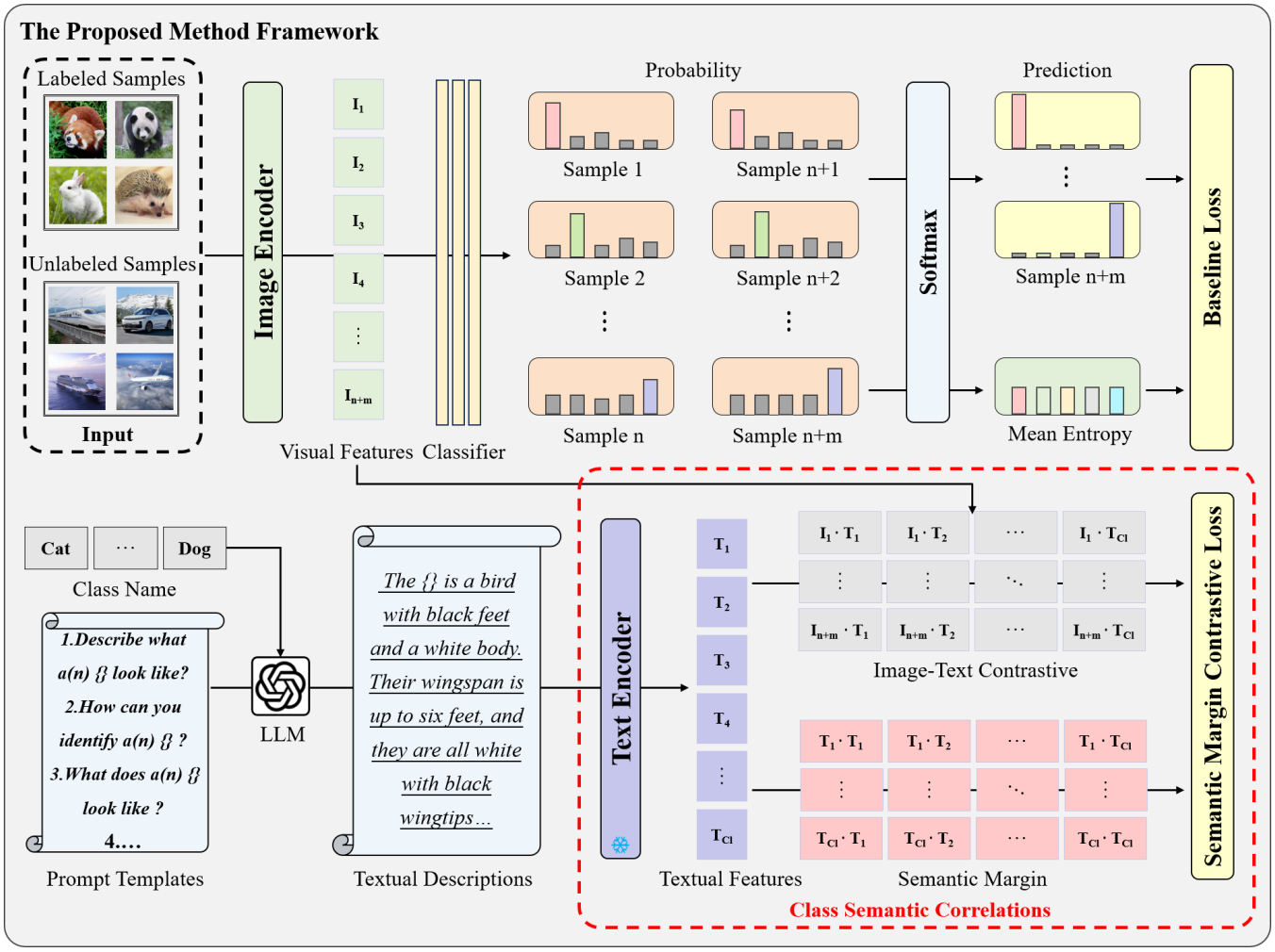

Figure 2: The proposed method framework. Our method fine-tunes only the image encoder during training while freezing the text encoder, thereby preserving the rich semantic correlations learned during the pre-training phase. Additionally, our method employs a semantic margin to extract class semantic correlations from textual descriptions, which are then used in enhancing image representation discriminability.

trastive loss is as follows:

$$L_{smc} = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\exp\left(\frac{S\left(I(x_i;\theta_I),T\left(t_{y_i};\theta_T\right)\right)}{\tau_k}\right)}{\sum_{k=1}^{|C|}\exp\left(\frac{S(I(x_i;\theta_I),T(t_k;\theta_T))+\beta M_{y_i,k}}{\tau_k}\right)} \tag{6}$$

where β is a hyper-parameter, which is set to 0.3 by default. Unlike L_{itc}, which treats all non-corresponding textual descriptions equally, L_{smc} takes into account the class semantic correlations. When the semantic similarity between classes y_i and k is small, indicating a distant semantic correlation, the value of the semantic margin $M_{y_i,k}$ is larger. This larger margin effectively distances the image features of x_i from the non-corresponding textual descriptions associated with class k. On the other hand, when the semantic similarity between y_i and k is significant, suggesting a closer semantic correlation, the semantic margin $M_{y_i,k}$ is reduced. This reduced margin allows the image features to be closer to the textual descriptions of classes that are semantically similar to y_i, thus enhancing the precision of distinguishing visual features from non-corresponding textual descriptions and facilitating the extraction of class semantic correlations.

Additionally, we refine the approach for obtaining textual descriptions. In the existing OWSSL methods, textual descriptions are typically acquired by querying the LLM directly, without sufficient consideration of the significance of prompt templates. It is recognized that effective prompt templates can significantly enhance the quality of the LLM's output. Therefore, we incorporate the prompt templates proposed in [Pratt *et al.*, 2023] to enhance the quality of the textual descriptions generated by the LLM.

Finally, the overall loss function of our proposed method is given as follows:

$$L = \lambda_1 L_{base} + \lambda_2 L_{smc}. \tag{7}$$

During the testing phase, the classification model $f(x;\theta_f)$ uses the classifier $h(I(x;\theta_I);\theta_h) : \mathbb{R}^d \rightarrow \mathbb{R}^{|\mathbb{C}|}$ to assign the

Dataset	Labeled		Unlabeled	
	# Image	# Class	# Image	# Class
CUB	1.5K	100	4.5K	200
Stanford Cars	2.0K	98	6.1K	196
Flowers102	0.3K	51	0.8K	102
ImageNet-100	31.9K	50	95.3K	100

Table 1: Statistics of the fine-grained datasets.

corresponding class labels to each input sample $x_i \in D_u$, thereby completing the classification of the samples.

4 Experiments

In this section, we conduct a comprehensive evaluation of our method. The experimental results and detailed analysis demonstrate the superiority of our method.

4.1 Experimental Setup

Datasets

We conduct experiments on five fine-grained datasets: CUB [Wah *et al.*, 2011] , Stanford Cars [Krause *et al.*, 2013] , Flowers102 [Nilsback and Zisserman, 2008], and ImageNet-100 [Deng *et al.*, 2009], which contain 200, 196, 102, and 100 classes, respectively. To ensure the fairness of the experiment, we conduct our experiments using the data partitioning method described in [Zheng *et al.*, 2024]. We adopt the same approach to divide the classes into known and unknown, considering 50% of the classes as known and the remaining 50% as unknown. Consequently, we construct the datasets D_l and D_u accordingly. We train our method on D_l and D_u, and subsequently evaluate its performance on D_u. This procedure is applied consistently across all compared methods. Detailed dataset information is provided in Table 1 .

Compared Methods

For OWSSL methods represent classes as symbolic variables, we select ORCA [Cao *et al.*, 2022], GCD [Vaze *et al.*, 2022], SimGCD [Wen *et al.*, 2023], GPC [Zhao *et al.*, 2023], DCCL [Pu *et al.*, 2023], and PromptCAL [Zhang *et al.*, 2023] as representative methods. For OWSSL methods that incorporate textual descriptions of classes, we choose CLIP-GCD [Ouldnoughi *et al.*, 2023] and TextGCD [Zheng *et al.*, 2024] as representative methods. These selected methods collectively showcase the latest advancements in the field of OWSSL.

Evaluation Protocol

Following the approach in [Cao *et al.*, 2022], we evaluate the performance of our method using accuracy (ACC) and clustering accuracy (CACC). Specifically, ACC and $CACC$ are calculated on dataset D_u, as illustrated in the following equations:

$$ACC = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbb{I}\left(\text{label}_i = \text{Result}_i\right), \tag{8}$$

$$CACC = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbb{I}\left(\text{label}_i = OP\left(\text{Result}_i\right)\right), \tag{9}$$

where label_i represents the ground-truth class label of $x_i \in D_u$, which is provided only during the testing phase. OP stands for optimal permutation, which aligns Result_i with label_i. Our evaluation includes calculating ACC for samples from known classes (K) in D_u, as well as calculating $CACC$ for all samples (A) and samples from unknown classes (U) in D_u.

Implementation Details

We use a ViT-B/16 [Dosovitskiy *et al.*, 2021] pre-trained with CLIP as the backbone of the image and text encoder, setting the output dimension of the backbone to 512. During training process, we adjust only the last block of the image encoder and freeze the text encoder. To ensure consistency with the experimental setup of other comparison methods, we follow the protocol described in TextGCD [Zheng *et al.*, 2024]. We use a batch size of 128 and train for 200 epochs with an initial learning rate of 0.1. We adjust the learning rate using a cosine annealing schedule. The parameters are set as follows: α to 0.35, λ_1 to 1, τ_s to 0.1, and τ_k to 0.01. τ_t is initialized to 0.07 and gradually increased to 0.04 using a cosine annealing schedule during the first 30 epochs of training. We choose GPT-3 [Brown *et al.*, 2020] as our LLM and utilize the prompt templates proposed by [Pratt *et al.*, 2023] to generate textual descriptions, and choose ViT-H based CLIP model as the auxiliary VLM. To validate the generalization capability of our method across different datasets, we pretrain the model using [Vaze *et al.*, 2022], uniformly setting λ_2 to 0.2 and β to 0.3. All our experiments are conducted on a single NVIDIA 3090 GPU.

4.2 Main Results

The classification accuracy of different methods across various datasets is provided in Table 2. We report the average maximum classification accuracy from three runs for our method, whereas the remaining results are obtained from [Wen *et al.*, 2023; Zheng *et al.*, 2024]. The experimental results show that our method significantly outperforms previous methods across various different datasets.

Compared to OWSSL methods that represent classes as symbolic variables, which can only indicate a sample's membership in a certain class and thus ignore the rich internal semantic information associated with the classes, this simplification causes the model to learn only a simple correspondence between samples and class symbols, resulting in weakly discriminative image representations, thereby limiting its ability to recognize unknown classes. Our method incorporates textual descriptions and aims to better understand the correspondence between the visual features of samples and the textual descriptions of classes. As a result, it outperforms methods that rely solely on symbolic variable representations.

Compared to OWSSL methods that incorporate textual descriptions of classes, these methods do not consider the class semantic correlations when aligning the visual features of samples with their corresponding textual descriptions, resulting in insufficiently discriminative image representations. Moreover, fine-tuning the text encoder may disrupt the rich semantic correlations learned during pre-training phase, po-

Methods	Pretrain	CUB			Stanford Cars			Flowers102			ImageNet-100		
		A	K	U	A	K	U	A	K	U	A	K	U
ORCA	DINO	35.5	45.6	30.2	23.5	50.1	10.7	-	-	-	73.5	92.6	63.9
GCD	DINO	51.3	56.6	48.7	39.0	57.6	29.9	74.4	74.9	74.1	74.1	89.8	66.3
SimGCD	DINO	60.3	65.6	57.7	53.8	71.9	45.0	71.3	80.9	66.5	83.0	93.1	77.9
GPC	DINO	55.4	58.2	53.1	42.8	59.2	32.8	-	-	-	76.9	94.3	71.0
DCCL	DINO	63.5	60.8	64.9	43.1	55.7	36.2	-	-	-	80.5	90.5	76.2
PromptCAL	DINO	62.9	64.4	62.1	50.2	70.1	40.6	-	-	-	83.1	92.7	78.3
GCD	CLIP	57.6	65.2	53.8	65.1	75.9	59.8	74.1	82.4	70.1	-	-	-
SimGCD	CLIP	62.0	76.8	54.6	75.9	81.4	73.1	75.3	87.8	69.0	86.1	94.5	81.9
CLIP-GCD	CLIP	62.8	77.1	55.7	70.6	88.2	62.2	76.3	88.6	70.2	84.0	**95.5**	78.2
TextGCD	CLIP	76.6	80.6	74.7	86.9	87.4	86.7	87.2	90.7	85.4	88.0	92.4	85.2
CSC-OWSSL	CLIP	**81.6**	**83.0**	**80.8**	**93.0**	**95.3**	**91.8**	**88.5**	**94.7**	**86.5**	**92.2**	94.7	**90.9**

Table 2: Classification accuracy (%) of compared methods on all, known, and unknown classes. Bold font indicates the best classification accuracy achieved on the corresponding dataset. A, K, and U denote model performance on all, known, and unknown classes, respectively.

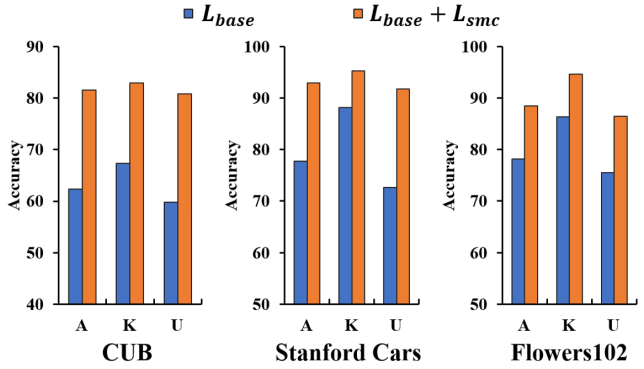

Figure 3: Ablation study on each component of the loss function on the different datasets.

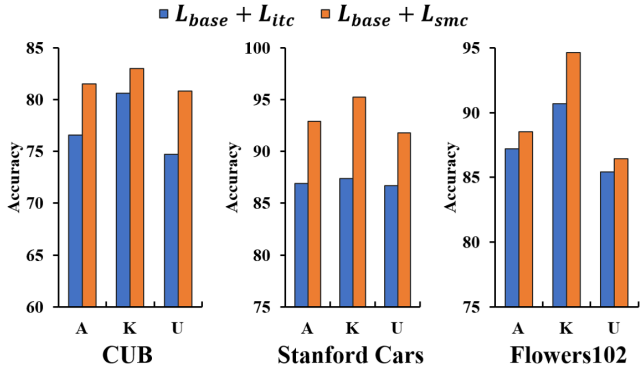

Figure 4: Performance comparison between the basic method and CSC-OWSSL.

tentially leading to overfitting to the provided textual descriptions. Both of these issues consequently affect the model's ability in recognizing unknown classes. In contrast, our method fine-tunes only the image encoder while keeping the text encoder frozen, thereby maintaining the semantic correlations established during pre-training. Additionally, we introduce a semantic margin contrastive loss to utilize class semantic correlations from textual descriptions. These combined strategies address the issues of the aforementioned methods and, as evidenced by our experimental results, significantly enhance performance.

4.3 Analyses and Discussions

Ablation Experiments

We conduct ablation studies on each component of the loss function. The results indicate that each component of the loss function contributes to the final performance. The specific results are presented in Figure 3.

Analysis of the Effect on Semantic Margin Contrastive Loss

To fully leverage the class semantic correlations, we freeze the text encoder to preserve the rich semantic correlations learned during the pre-training phase and employ a semantic margin to extract class semantic correlations from textual descriptions. This strategy allows us to utilize the class semantic correlations from textual descriptions, thereby effectively enhancing image representation discriminability and the model's ability to recognize unknown classes. To demonstrate the effectiveness of our strategy, we conduct experiments comparing a basic method that uses the loss function $L_{base}+L_{itc}$ and fine-tunes the text encoder (which essentially is TextGCD [Zheng *et al.*, 2024]) with our method, which employs $L_{base} + L_{smc}$. The experimental results, presented in Figure 4, show that our strategy significantly outperforms the basic method.

Analysis of the Sensitivity of Hyper-Parameters

We introduce an additional hyper-parameter β into the L_{smc}, while all other hyper-parameters remain consistent with the references and are maintained at the same values across all experiments, as detailed in the experimental section. Sensitivity analyses for β are presented in Figure 5. These analyses demonstrate that the performance of our method is robust to the choice of the hyper-parameter β.

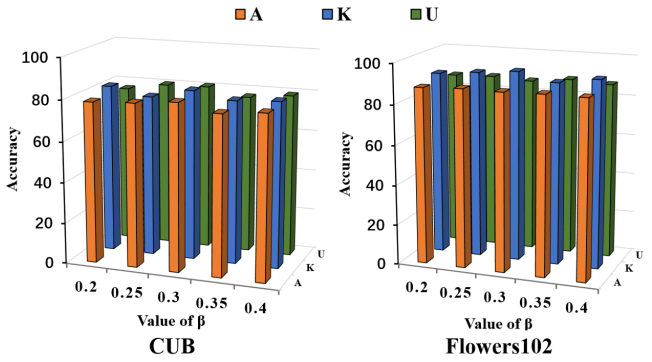

Figure 5: Performance of CSC-OWSSL with different values of hyper-parameters.

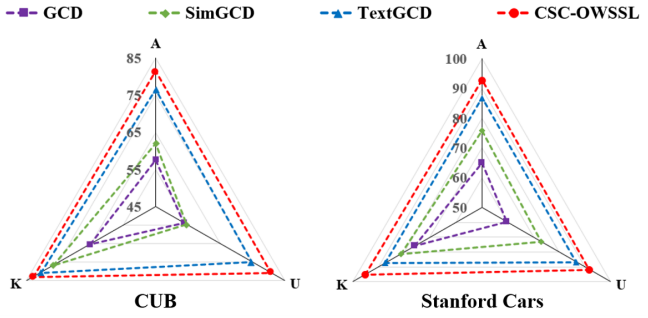

Figure 6: Performance improvement of semantic margin for OWSSL frameworks.

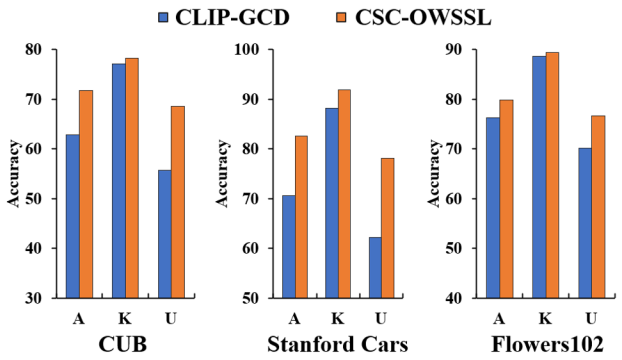

Figure 7: Performance of CSC-OWSSL with textual description of known classes.

Analysis of the Performance Improvement of Semantic Margin for OWSSL Frameworks

In the field of OWSSL, several frameworks are incorporated within the existing methods. For instance, GCD [Vaze *et al.*, 2022] constructs a discriminative representation space through semi-supervised contrastive learning, and SimGCD [Wen *et al.*, 2023] enhances performance via a self-distillation strategy. Most subsequent methods are improvements based on these two frameworks [Zhao *et al.*, 2023; Pu *et al.*, 2023; Zhang *et al.*, 2023]. However, these frameworks commonly simplify classes into symbolic variables, which can determine the classification of samples but ignore the rich semantic information of the classes. Therefore, TextGCD [Zheng *et al.*, 2024] attempts to address this issue by incorporating textual information and employing an image-text contrastive loss to align visual features with textual descriptions. As a result, it is considered an OWSSL framework. Nevertheless, it still fails to consider the class semantic correlations. Our method keeps the text encoder frozen, thus preserving the rich semantic correlations learned during the pre-training phase, and introduces a semantic margin contrastive loss to utilize class semantic correlations from textual descriptions. These combined strategies maintain the semantic correlations established during pre-training and collectively address the issues of these frameworks. Across multiple experimental metrics, our method demonstrates performance that surpasses the current frameworks, with the comparative experimental results detailed in Figure 6 .

Analysis of Performance in More Realistic Scenarios

Existing OWSSL methods that utilize textual descriptions typically assume that the names of all classes are pre-known. However, this assumption often does not hold in real-world scenarios, as in practical applications, we usually only have access to information about a subset of classes. To address this shortcoming, in this section, we investigate a scenario that is closer to reality: validating the performance of our proposed method using only the textual descriptions of known classes. Furthermore, we do not employ any auxiliary VLMs, which might not be available or practical in real-world settings. All other conditions remain the same. This allows us to more accurately evaluate the generalization capability of our method when dealing with real-world problems. The relevant experimental results are presented in Figure 7. The results show that even when using only the textual descriptions of known classes, our method still demonstrates competitive performance compared to OWSSL methods that use textual descriptions for all classes.

5 Conclusion

In this paper, we propose a novel OWSSL method. This method fine-tunes only the image encoder during training while keeping the text encoder frozen, thereby preserving the rich semantic correlations learned during the pre-training phase. Furthermore, our method introduces a semantic margin to extract class semantic correlations from textual descriptions. These semantic correlations are then utilized to enhance the discriminability of image representations. By focusing on these aspects, our method significantly improves model performance compared to representative OWSSL methods. Future research aims to explore and address the challenges posed by the continuous emergence of unknown classes, a scenario commonly encountered in practical applications, while investigating the capabilities of our method in such environments.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62376141, U21A20473, 62376142, 62276158).

References

[An *et al.*, 2024] Wenbin An, Feng Tian, Wenkai Shi, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. Transfer and alignment network for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10856–10864, 2024.

[Assran *et al.*, 2022] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Proceedings of the European Conference on Computer Vision*, pages 456–473, 2022.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1877–1901, 2020.

[Cao *et al.*, 2022] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

[Han *et al.*, 2019] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8400–8408, 2019.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Li *et al.*, 2021] YuFeng Li, LanZhe Guo, and ZhiHua Zhou. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, 43(1):334–346, 2021.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[Oliver *et al.*, 2018] Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3239–3250, 2018.

[Ouldnoughi *et al.*, 2023] Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. CLIP-GCD: simple language guided generalized category discovery. *CoRR*, abs/2305.10420, 2023.

[Pratt *et al.*, 2023] Sarah M. Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero- shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15645–15655, 2023.

[Pu *et al.*, 2023] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7588, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[Riz *et al.*, 2023] Luigi Riz, Cristiano Saltori, Elisa Ricci, and Fabio Poiesi. Novel class discovery for 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9393–9402, 2023.

[Rizve *et al.*, 2022] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *Proceedings of the European Conference on Computer Vision*, pages 382–401, 2022.

[Scheirer *et al.*, 2013] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[Shi *et al.*, 2024] Wenkai Shi, Wenbin An, Feng Tian, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. A unified knowledge transfer network for generalized category discovery. In Michael J. Wooldridge, Jennifer G. Dy,

and Sriraam Natarajan, editors, *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18961–18969, 2024.

[Shu *et al.*, 2023] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing CLIP to out-of-distributions. In *Proceedings of the International Conference on Machine Learning*, pages 31716–31731, 2023.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 596–608, 2020.

[Tu *et al.*, 2024] Yuanpeng Tu, Yuxi Li, Boshen Zhang, Liang Liu, Jiangning Zhang, Yabiao Wang, and Cairong Zhao. Self-supervised likelihood estimation with energy guidance for anomaly segmentation in urban scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21637–21645, 2024.

[Vaze *et al.*, 2022] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2022.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

[Wen *et al.*, 2023] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16544–16554, 2023.

[Zang *et al.*, 2024] Yuan Zang, Tian Yun, Hao Tan, Trung Bui, and Chen Sun. Pre-trained vision-language models learn discoverable visual concepts. *CoRR*, abs/2404.12652, 2024.

[Zhang *et al.*, 2023] Sheng Zhang, Salman H. Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488, 2023.

[Zhao *et al.*, 2023] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16577–16587, 2023.

[Zheng *et al.*, 2024] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *Proceedings of the European Conference on Computer Vision*, pages 41–58, 2024.