Debiased Textual Prompt Tuning for Enhancing Unknown Class Discovery

Yuxin Fan[1], Junbiao Cui[1,⋆], Changhao Liu[1], Xingwang Zhao[1], and Jiye Liang[1]

[1] 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,cjb,liuchanghao,zhaoxw,ljy}@sxu.edu.cn

Abstract. Open-world semi-supervised learning (OWSSL) aims to recognize known and unknown classes from unlabeled samples, where labeled samples only include known classes. Recent studies introduce class-specific textual descriptions as semantic information and employ learnable textual prompts to optimize these descriptions, thereby enhancing the model's ability to recognize classes. However, these methods often overlook the model bias toward known classes during textual prompt tuning, and establish only limited cross-modal connections, leading to severe sparsity in the graph structure and hindering the propagation of visual and semantic information across the graph. These issues collectively limit the model's ability to discover unknown classes, particularly in scenarios with a large number of classes. To address these issues, we propose a novel OWSSL method, which consists of two core components: (1) textual prompt optimization aimed at mitigating the model bias toward known classes, and (2) bimodal graph enhancement designed to alleviate the sparsity issue in the graph structure. Experimental results across multiple datasets indicate that our method achieves an average accuracy 20.7% higher on unknown classes than state-of-the-art OWSSL methods (relative improvement).

Keywords: Open-World Learning · Unknown Class Discovery · Textual Prompt Tuning

1 Introduction

Machine learning [23], particularly supervised learning, has achieved breakthroughs across various domains [10, 11], yet high annotation costs restrict its broader adoption. To mitigate this, semi-supervised learning (SSL) [12, 41] has been introduced, leveraging limited labeled samples and abundant unlabeled samples for training, thereby broadening the application domain. Traditional SSL methods achieve significant advancements in closed-world scenarios where all classes are represented in the labeled samples [50], but they fall short in open-world scenarios with unlabeled samples containing potential unknown classes. This shortcoming stems from their design paradigm, which lacks an intrinsic mechanism

⋆ Corresponding author

for discovering unknown classes [13, 25]. To address this challenge, open-world semi-supervised learning (OWSSL) is proposed [1, 45], which aims to effectively recognize both known and unknown classes in the unlabeled samples when labeled samples only include the known classes. Figure 1 illustrates the scenario encountered by OWSSL.

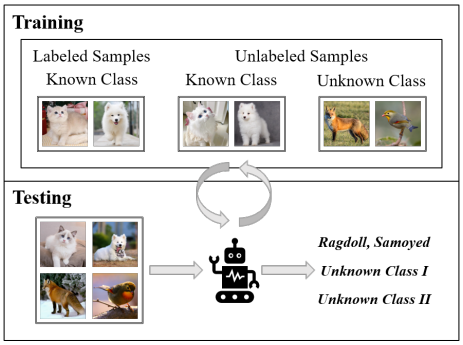

Fig. 1: The environment faced by OWSSL. The model needs to be able to effectively recognize both known and unknown classes after training.

Existing OWSSL methods primarily represent classes as one-hot class labels, which merely indicate the class-specific membership of samples [37, 48]. This symbolic representation ignores the rich semantic information associated with the classes, limiting the model's ability to discover unknown classes. In response to this issue, recent studies [9, 51] introduce class-specific textual descriptions as semantic information and employ a CLIP-inspired image-text contrastive learning framework [32, 36] for training. However, these textual descriptions introduce background noise irrelevant to the classification task, potentially diminishing the model's effectiveness in image-text alignment. Therefore, the latest studies [8, 49] further incorporate learnable textual prompts to optimize these descriptions based on visual features, thereby effectively enhancing the model's ability to recognize classes.

However, two major issues limit the effectiveness of these text-guided OWSSL methods in discovering unknown classes: (1) They often overlook model bias toward known classes during textual prompt tuning, which arises from the lack of supervision for unknown classes [1]. This bias inclines the model to prefer classifying samples as known classes, which suppresses the optimization process for learning effective textual representations of unknown classes, thereby preventing the adequate capture of their discriminative features. (2) They often establish only limited cross-modal connections. Since both visual features and textual descriptions can be regarded as graph nodes from graph machine learning [42], the resulting graph structure suffers from sparsity issues, thereby hindering the propagation of visual and semantic information across the graph [52]. These issues collectively limit the model's ability to discover unknown classes, particularly in scenarios with a large number of classes.

To address these issues, we propose a novel OWSSL method comprising two core components: (1) Textual Prompt Optimization: This aims to mitigate the model bias towards known classes during tuning [5, 35], thus enabling the optimized prompts to generate less biased and more effective textual representations for unknown classes, thereby enhancing the model's ability to capture discriminative features. (2) Bimodal Graph Enhancement: This involves constructing a proximity graph for image-image and image-prompt pairs of unlabeled samples

to alleviate the sparsity issue in the graph structure. Additionally, bimodal label propagation [42] is applied to facilitate the effective propagation of visual and semantic information across the graph. Experimental results on multiple datasets demonstrate that our method significantly outperforms state-of-the-art OWSSL methods across multiple metrics. To summarize, the main contributions of this paper are as follows:

- We identify two key issues in existing text-guided OWSSL methods: they often overlook model bias toward known classes during textual prompt tuning, and establish only limited cross-modal connections, which results in sparse graph structures. These issues limit the model's ability to discover unknown classes.
- We propose a novel OWSSL method comprising two core components: textual prompt optimization that generates less biased and more effective representations, and bimodal graph enhancement that alleviates graph sparsity to facilitate effective propagation of visual and semantic information across the graph.
- We conduct comprehensive experiments on multiple datasets with a large number of classes. Experimental results demonstrate that our method outperforms state-of-the-art OWSSL methods on unknown classes by an average relative accuracy improvement of 20.7%.

2 Related Work

2.1 Open-World Machine Learning

Most machine learning methods operate under the closed-world assumption, but this assumption often proves inadequate in real-world scenarios. To address this limitation, the field of open-world machine learning has emerged, encompassing research directions such as open-set recognition (OSR) [38], robust semi-supervised learning (Robust SSL) [31], and novel class discovery (NCD) [15], all aimed at adapting to more open environments. OSR mandates that models identify unknown classes without hindering performance on known classes. Robust SSL posits that unlabeled samples include classes not present in the labeled samples, with the goal of minimizing the adverse effects of these unknown classes on model performance for known classes. NCD assumes that unlabeled samples are composed exclusively of unknown classes, necessitating the model to uncover these unknown classes. Although these research directions have made some progress, they do not fully encompass the complexity of open-world scenarios. To better adapt to the complexities of the real-world, OWSSL is proposed to extend the above-mentioned environment, aiming to contribute to the advancement of open-world machine learning. This research setting is also known as generalized category discovery.

2.2 Open-World Semi-Supervised Learning

OWSSL is currently receiving significant attention, with numerous methods being proposed. GCD [45] constructs a discriminative representation space through semi-supervised contrastive learning; SimGCD [48] builds on this foundation and enhances model performance with a self-distillation strategy. Subsequent studies, such as DCCL [34] and SelEx [37], further improve upon these methods. However, these methods primarily represent classes using one-hot class labels, which merely indicate the class-specific membership of samples. Therefore, TextGCD [51] integrates class-specific textual descriptions as semantic information and introduces a retrieval-based text generation strategy. Furthermore, TP-OWSSL [8] incorporates learnable textual prompts to optimize these descriptions based on visual features, thereby enhancing the model's ability to recognize classes. In addition, CSC-OWSSL [9] utilizes class semantic correlations to enhance image representation discriminability and becomes the state-of-the-art method in the field of OWSSL. However, these text-guided methods often overlook the model bias towards known classes and establish only limited cross-modal connections. This restricts the model's ability to discover unknown classes, particularly in scenarios with a large number of classes. Therefore, this paper conducts experiments on datasets with a large number of classes to evaluate the model performance under such challenging conditions.

2.3 Open-World Class Discovery in Diverse Scenarios

This paper focuses on the classic text-guided OWSSL setting. Meanwhile, to further advance open-world unknown class discovery, recent studies explore various challenging and realistic extended settings. For instance, Continual Category Discovery (CCD) [20, 29] aims to address incremental learning over time; Federated Category Discovery (FCD) [33] focuses on decentralized model training under privacy constraints; and Semantic Category Discovery (SCD) [14] is dedicated to assigning semantic labels from an open vocabulary to unlabeled samples. Additionally, other research efforts tackle scenarios such as few-shot learning [4], long-tailed distributions [26], and domain shifts [47], collectively enhancing the practicality and robustness of these methods in real-world applications. In response to these diverse research advancements, recent surveys [17, 43, 53] systematically review and summarize these extended settings and their associated literature, thereby offering structured insights into this research field.

3 Proposed Method

In this section, we first describe the OWSSL setting. We then introduce our proposed method TPOBGE, which consists of two components: Textual Prompt Optimization and Bimodal Graph Enhancement. The overall framework of the proposed method is depicted in Figure 2 and the detailed algorithm is presented in Algorithm 1.

3.1 Problem Setting

Given an unlabeled dataset D_u with m samples, denoted as $D_u = \{x_1, x_2, \ldots, x_m\}$, and a labeled dataset D_l with n samples, denoted as $D_l = \{(x_{m+1}, y_{m+1}), (x_{m+2}, y_{m+2}), \ldots, (x_{m+n}, y_{m+n})\}$, it is typically assumed that $m \gg n$, where x_i represents a sample and y_i represents the corresponding one-hot class label. C_u denotes the set of classes present in the unlabeled dataset D_u, while C_l denotes the set of classes present in the labeled dataset D_l. In OWSSL settings, $C_{known} = C_l$ is the set of known classes and $C_{unknown} = C_u \setminus C_l$ is the set of unknown classes. OWSSL methods aim to train a classification model capable of assigning samples of unknown classes to both known and unknown classes. The classification model is trained on D_l and D_u, and then evaluated on D_u. Existing OWSSL methods generally assume that the number of unknown classes is known, and our method aligns with this assumption. Moreover, in scenarios where the number of unknown classes is unclear, the approach introduced by GCD [45] to estimate the number of unknown classes can be employed.

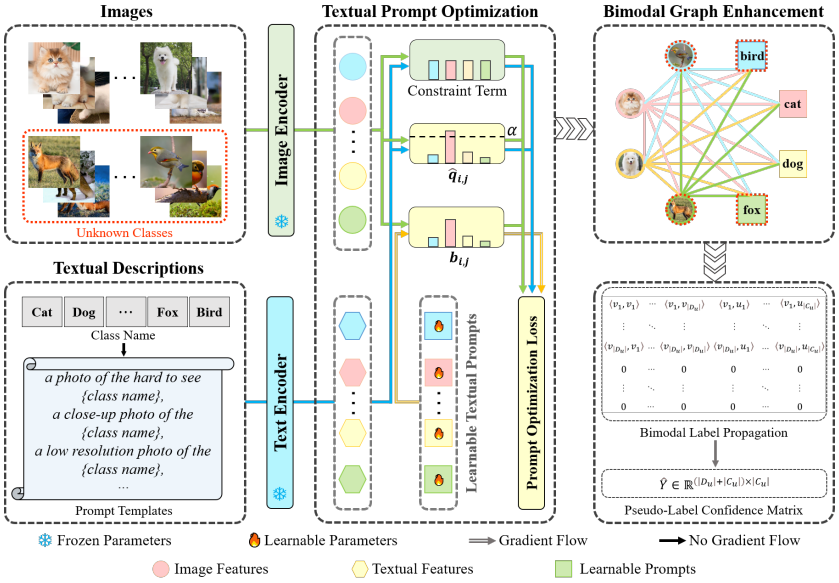

Fig. 2: The overall framework of TPOBGE consists of textual prompt optimization and bimodal graph enhancement. This effectively generates less biased textual representations of unknown classes and alleviates the sparsity issue in the graph structure.

In this paper, we construct the classification model $f(x; \theta_f)$ using deep neural networks. The model architecture comprises an image encoder $I(x; \theta_I) : \mathbb{R}^{D_I} \to \mathbb{R}^d$ and a text encoder $T(t; \theta_T) : \mathbb{R}^{D_T} \to \mathbb{R}^d$, where x represents an image, t denotes textual descriptions, and $\mathbb{R}^{D_I}$ and $\mathbb{R}^{D_T}$ are the input dimensions for the image and text encoders, respectively.

Algorithm 1 Detailed algorithm of TPOBGE.

Input: Labeled dataset D_l, unlabeled dataset D_u, number of classes in unlabeled dataset $|C_u|$ (which could be ground truth or estimated), class description database $ClassDescription = \{t_i\}_{i=1}^{K}$.

Input: Image encoder $I(x;\theta_I)$, text encoder $T(t;\theta_T)$, learnable prompts $Prompt = \{u_i\}_{i=1}^{|C_u|}$.

Input: Hyperparameters τ_T, τ_I, k, β.

1: Extract image features and textual features from datasets and descriptions using $I(x;\theta_I)$ and $T(t;\theta_T)$.
2: Alleviate model bias using Eqn. (4).
3: Convert high-confidence data into one-hot pseudo-labels using Eqn. (5).
4: Optimize the textual prompts using Eqn. (6).
5: Construct a bimodal proximity graph P.
6: Apply label propagation using Eqn. (7) and (8).
7: **return** Optimized pseudo-label confidence matrix $\hat{Y}$.

Output: For $x_i \in D_u$, its predicted class label is $\hat{y}_i = \arg\max_c \hat{Y}(i,c)$.

3.2 Baseline

Existing text-guided OWSSL methods typically introduce class-specific textual descriptions as semantic information, a framework our method also follows [9,51]. Moreover, it is worth noting that including descriptions related to unknown classes is a common and widely adopted practice in open-world machine learning, a convention our method adheres to. Therefore, we adopt a similarly textual description acquisition strategy. Specifically, the process begins by constructing a comprehensive description database, denoted as $ClassDescription = \{t_i\}_{i=1}^{K}$. Subsequently, the similarity between each unlabeled image $x_i \in D_u$ and every candidate description $t_i \in ClassDescription$ is computed using the formula $\langle I(x_i;\theta_I), T(t_i;\theta_T)\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the dot product. For each image $x_i \in D_u$, it is assigned to the description t_i for which their similarity is highest. Finally, the frequency of each assigned description is counted, the top $|C_u|$ most frequent ones are selected, and the refined set $ClassDescription_u = \{t_i\}_{i=1}^{|C_u|}$ is formed. The value of $|C_u|$ could be ground truth or estimated.

However, these textual descriptions introduce background noise irrelevant to the classification task [22], potentially diminishing the model's effectiveness in image-text alignment. Therefore, the latest studies [8, 49] further incorporate learnable textual prompts to optimize these descriptions based on visual features, thereby enhancing the model's ability to recognize classes. Following this paradigm, we construct learnable prompts for each class, which are randomly initialized vectors, denoted as $Prompt = \{u_i\}_{i=1}^{|C_u|}, u_i \in \mathbb{R}^d$, which are optimized as follows:

$$L = \sum_{i=1}^{|D_u|+|D_l|} \sum_{j=1}^{|C_u|} q_{i,j} \log\left(\frac{q_{i,j}}{b_{i,j}}\right), \tag{1}$$

where $q_{i,j}$ and $b_{i,j}$ are defined as:

$$q_{i,j} = \frac{\exp\left(\langle I(x_i;\theta_I), T(t_j;\theta_T)\rangle / \tau_T\right)}{\sum_{k=1}^{|C_u|} \exp\left(\langle I(x_i;\theta_I), T(t_k;\theta_T)\rangle / \tau_T\right)}, \tag{2}$$

$$b_{i,j} = \frac{\exp\left(\langle I(x_i;\theta_I), u_j\rangle / \tau_I\right)}{\sum_{k=1}^{|C_u|} \exp\left(\langle I(x_i;\theta_I), u_k\rangle / \tau_I\right)}. \tag{3}$$

with τ_T and τ_I being temperature hyper-parameters, $\langle\cdot,\cdot\rangle$ being dot product.

3.3 Textual Prompt Optimization

However, the aforementioned process overlooks the model bias towards known classes during prompt tuning, which suppresses the optimization process for learning effective textual representations of unknown classes, thereby preventing the adequate capture of their discriminative features. To address this, we impose constraints on this process to alleviate model bias during prompt tuning [5, 35]. The associated loss function is as follows:

$$\begin{aligned} \tilde{Q} = \underset{Q}{\operatorname{argmax}}\ & \langle Q, S\rangle + \tau_T H(Q) \\ \text{s.t.}\ & \begin{cases} \sum_{j=1}^{|C_u|} q_{i,j} = \frac{1}{|D_u|+|D_l|}, \forall i \in \{1,\ldots,|D_u|+|D_l|\}, \\ \sum_{i=1}^{|D_u|+|D_l|} q_{i,j} = \frac{1}{|C_u|}, \forall j \in \{1,\ldots,|C_u|\}, \\ q_{i,j} \geq 0, \forall i \in \{1,\ldots,|D_u|+|D_l|\}, \forall j \in \{1,\ldots,|C_u|\}. \end{cases} \end{aligned} \tag{4}$$

with $Q = [q_{i,j}]_{i=1,\cdots,|D_u|+|D_l|}^{j=1,\cdots,|C_u|}$, $S = [s_{i,j}]_{i=1,\cdots,|D_u|+|D_l|}^{j=1,\cdots,|C_u|}$, $s_{i,j} = \langle I(x_i;\theta_I), T(t_j;\theta_T)\rangle$, and $H(Q)$ computes the entropy of Q. The first constraint is similar to probability normalization, which ensures that the sum of the probability distribution for each data equals a constant. The second constraint is a class correction. It mechanically ensures a uniform distribution of assignments across classes, thereby preventing the model from developing a bias and allocating fair predictive probability to unknown classes. In essence, this regularization mechanism can be viewed as solving an optimal transport (OT) problem. We employ the Sinkhorn algorithm [5] to solve the entropy-regularized OT problem, with the marginal power scaling parameter set to 0 and the number of iterations set to 20. Since the data distribution is unknown, we adopt the standard OT setting by default, which ensures good performance across different data distributions. If the data distribution is known, we can employ the unbalanced OT method [39, 40].

Moreover, following semi-supervised learning principles [41], we convert high-confidence data into one-hot pseudo-labels to mitigate the impact of irrelevant classes. Specifically, we normalize the $\tilde{Q}$ as $\tilde{Q}^{norm} = (|D_u| + |D_l|) \cdot \tilde{Q}$, and it is further optimized as:

$$\hat{Q}_i = \begin{cases} e_{j^*}, & j^* = \underset{j \in \{1,\cdots,|C_u|\}}{\arg\max} \{\tilde{q}_{i,j}^{norm}\}, \tilde{q}_{i,j^*}^{norm} > \alpha, \\ \tilde{Q}_i^{norm}, \text{otherwise.} \end{cases} \tag{5}$$

where $\tilde{Q}_i^{norm} = \left[\tilde{q}_{i,j}^{norm}\right]_{j=1}^{|C_u|}$, $e_{j^*} \in \{0,1\}^{|C_u|}$ is a one-hot vector with the j^*-th element being 1, and α is a predefined threshold set to 0.5. Finally, the loss function of prompt optimization is as follows:

$$L = \sum_{i=1}^{|D_u|+|D_l|} \sum_{j=1}^{|C_u|} \hat{q}_{i,j} \log\left(\frac{\hat{q}_{i,j}}{b_{i,j}}\right). \tag{6}$$

3.4 Bimodal Graph Enhancement

Construct a Bimodal Proximity Graph. Existing text-guided OWSSL methods mainly focus on the similarity between the visual features of samples and their corresponding textual descriptions, thereby establishing only limited cross-modal connections [8, 9]. From the perspective of graph machine learning, both visual features and textual descriptions can be regarded as nodes [42], which results in intra-modal sparsity (a lack of connections between the visual features of different samples) and inter-modal sparsity (sparse connections between the visual features of samples and textual descriptions of multiple classes), ultimately leading to severe sparsity within the graph structure. This severe sparsity hinders the propagation of visual and semantic information within the graph [52], thereby limiting the model's ability to discover unknown classes.

To address this issue, we construct a bimodal proximity graph, utilizing the intra-modal and inter-modal similarities between image-image and image-prompt pairs of unlabeled samples, thus establishing intra-modal and inter-modal edges and effectively mitigating the sparsity issues of the graph structure. Specifically, the steps of constructing the bi-modal proximity graph are as follows:

(1) We define the visual feature set of the unlabeled samples $Vision = \{v_i = I(x_i; \theta_I)\}_{i=1}^{|D_u|}$ and introduce the prompt set of all classes $Prompt = \{u_i\}_{i=1}^{|C_u|}$.

(2) For the sample $x_i \in D_u$, we first calculate the similarity between v_i and each element in the set $Vision$, and record the results as $S_{image} = \{\langle v_i, v_j \rangle\}_{j=1}^{|D_u|}$. Subsequently, we calculate the similarity between v_i and each element in the set $Prompt$, and record the results as $S_{text} = \{\langle v_i, u_j \rangle\}_{j=1}^{|C_u|}$. To reduce the complexity of the graph, we retain the top k elements from S_{image} and S_{text} respectively, and set the remaining elements to 0. Specifically, in S_{image}, the element $\langle v_i, v_i \rangle$ is not considered among the top k elements, where k is a hyperparameter. Then, we concatenate the elements of S_{image} and S_{text}. Ultimately, for the sample $x_i \in D_u$, we obtain a vector $p_i \in \mathbb{R}^{(|D_u|+|C_u|)}$.

(3) We calculate p_i for each sample $x_i \in D_u$, and then concatenate all p_i along the row direction to obtain a proximity matrix $P_{image} \in \mathbb{R}^{|D_u| \times (|D_u|+|C_u|)}$.

(4) For each textual description t_i, we follow the calculation methods of steps (2) and (3) to obtain $P_{text} \in \mathbb{R}^{|C_u| \times (|D_u|+|C_u|)}$. Then, we concatenate

P_{image} and P_{text} along the row direction to obtain the bi-modal proximity matrix $P \in \mathbb{R}^{(|D_u|+|C_u|)\times(|D_u|+|C_u|)}$.

Finally, we construct a bi-modal proximity graph $G = (Nodes, Edges)$, where $Nodes = [Vision; Prompt]$ represents the concatenated set of nodes, and $Edges = P$ represents the set of edges.

Apply Bimodal Label Propagation. After constructing the bi-modal proximity graph, we need to apply the label propagation algorithm to facilitate the effective propagation of visual and semantic information across the graph [42], thereby enhancing the model's ability to recognize classes. However, the performance of label propagation algorithms is heavily dependent on the quality of the graph structure. Research indicates a significant modality gap between the image and text representations from vision-language models [27, 44], meaning that visual and textual similarities are often inconsistent. This similarity mismatch prevents the constructed proximity graph from reliably reflecting true semantic relationships, consequently compromising the effectiveness of label propagation.

Therefore, when constructing the bi-modal proximity graph, we avoid using the similarity between textual features and instead set all elements of P_{text} to zero, obtaining the updated $\tilde{P}$ and $\tilde{G}$. The steps for label propagation on the updated bi-modal proximity graph are as follows:

(1) We need to initialize the pseudo-labels for each node in graph $\tilde{G}$. For this purpose, we construct a pseudo-label matrix $Y \in \mathbb{R}^{(|D_u|+|C_u|)\times|C_u|}$. For each visual feature node $v_i \in Vision$ lacking pseudo-labels, the first $|D_u|$ rows of Y are initialized to zero. For each textual feature node $u_i \in Prompt$, its corresponding pseudo-label is represented in Y as follows:

$$Y(|D_u| + i, c) = \begin{cases} 1, & \text{if } u_i \text{ is textual feature of class } c \\ 0, & \text{otherwise} \end{cases}. \tag{7}$$

(2) We normalize $\tilde{P}$ to obtain $\hat{P}$, where $\hat{P} = D^{-\frac{1}{2}}\left(\tilde{P} + \tilde{P}^T\right)D^{-\frac{1}{2}}$, and D is the degree matrix of $(\tilde{P} + \tilde{P}^T)$. We then use the label propagation algorithm [52] to iteratively update the pseudo-label confidence matrix $\hat{Y} \in \mathbb{R}^{(|D_u|+|C_u|)\times|C_u|}$, thereby facilitating the effective propagation of visual and semantic information across the graph. The update formula is as follows:

$$\hat{Y}^{(r+1)} = \beta\hat{P}\hat{Y}^{(r)} + (1-\beta)\,Y. \tag{8}$$

Here, β is a hyperparameter that controls the propagation weight, and r is the current iteration. The initial value is set as $\hat{Y}^{(0)} = Y$. The efficient implementation of the label propagation algorithm is thoroughly discussed in existing literature [18, 19], and thus we do not discuss it further.

3.5 Inference

During the testing phase, $f(x;\theta_f)$ classifies each input sample $x_i \in D_u$. For $x_i \in D_u$, its predicted class label is $\hat{y}_i = \arg\max_c \hat{Y}\,(i, c)$.

4 Experiments

In this section, we conduct a comprehensive evaluation of our method. The experimental results and detailed analyses demonstrate its superiority over existing OWSSL methods.

4.1 Experimental Setup

Datasets. We conduct comprehensive experiments on four datasets with a large number of classes: Hybrid, ImageNet-500, ImageNet-1K [6], and Webvision [24], which contain 498, 500, 1k, and 1k classes, respectively. Hybrid is composed of CUB [46], Stanford Cars [21], and Flowers [30] datasets. For all datasets, we consider the first 50% of classes as known classes and the remaining 50% as unknown classes. We select 50% of the samples from known classes to construct D_l, with the rest serving as D_u. 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. Dataset information is provided in Table 1.

Table 1: Statistics of datasets.

Dataset	Labeled		Unlabeled	
	Image	Class	Image	Class
Hybrid	3.7k	249	11.4k	498
ImageNet-500	6.2k	250	18.7k	500
ImageNet-1K	12.5k	500	37.5k	1k
Webvision	12.5k	500	37.5k	1k

In the construction of the Hybrid dataset, we apply the CSC-OWSSL [9] partitioning approach to the CUB, Stanford Cars, and Flowers datasets. This involves first dividing each dataset into known and unknown classes, and subsequently aggregating the known classes from all three to form the known classes of Hybrid, and similarly aggregating the unknown classes to form its unknown classes. The ImageNet-500 dataset is constructed by selecting the top 500 classes from ImageNet-1K. For the WebVision dataset, we choose version V1. Additionally, to reduce the resource consumption during training, we use the validation sets of ImageNet-1K and WebVision to construct these datasets.

Compared Methods. For OWSSL methods represent classes as one-hot class labels, we select GCD [45], SimGCD [48], LegoGCD [2], and ProtoGCD [28] as representative methods. For OWSSL methods that incorporate textual descriptions of classes, we choose TP-OWSSL [8], CSC-OWSSL [9], SSR2-GCD [16] and SpectralGCD [3] as representative methods. These selected methods showcase the representative and latest advancements in the field of OWSSL.

Table 2: Classification accuracy (%) of compared methods. 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.

Methods	Pretrain	Hybrid			ImageNet-500			ImageNet-1K			Webvision			Average		
		A	K	U	A	K	U	A	K	U	A	K	U	A	K	U
GCD	DINO	39.45	48.28	30.37	58.89	65.71	55.47	51.25	56.06	48.84	45.10	50.59	42.35	48.67	55.16	44.26
SimGCD	DINO	49.35	55.73	42.78	45.60	66.56	35.12	34.34	60.84	21.09	31.80	56.67	19.36	40.27	59.95	29.59
LegoGCD	DINO	48.30	52.64	43.85	46.49	69.81	34.82	34.20	61.13	20.74	31.60	56.85	18.97	40.15	60.11	29.60
ProtoGCD	DINO	39.68	46.84	32.27	33.59	63.23	18.77	31.28	63.05	15.40	31.57	56.80	18.96	34.03	57.48	21.35
GCD	CLIP	52.95	54.30	51.56	51.26	64.22	44.78	43.82	53.55	38.95	41.41	50.13	37.05	47.36	55.55	43.08
SimGCD	CLIP	64.62	64.14	65.11	54.73	**76.56**	43.82	36.54	61.73	23.94	34.93	58.59	23.11	47.70	65.25	38.99
LegoGCD	CLIP	63.45	64.40	62.46	54.82	72.75	45.85	38.00	63.17	25.41	36.00	58.50	24.75	48.06	64.70	39.61
ProtoGCD	CLIP	48.19	50.39	45.93	34.58	65.15	19.30	32.22	62.53	17.07	30.15	57.29	16.58	36.28	58.84	24.72
TP-OWSSL	CLIP	44.81	42.78	45.15	68.98	71.71	67.69	-	-	-	-	-	-	-	-	-
CSC-OWSSL	CLIP	62.91	**68.35**	60.19	56.74	72.13	49.04	39.03	63.05	27.02	34.94	58.22	23.29	48.41	65.44	39.89
SSR2-GCD	CLIP	55.70	58.71	52.60	65.41	70.54	62.84	58.73	60.02	58.08	55.42	58.43	53.92	58.81	61.92	56.86
SpectralGCD	CLIP	63.79	63.47	64.15	56.49	71.96	48.75	64.01	**78.53**	56.74	53.92	**68.74**	46.50	59.55	**70.67**	54.03
TPOBGE	CLIP	**63.81**	62.40	**65.27**	**73.41**	73.47	**73.40**	**69.70**	71.39	**68.98**	**66.73**	66.63	**66.96**	**68.41**	68.47	**68.65**

Evaluation Protocol. Following [45], we use the clustering accuracy (ACC) to evaluate the performance of our method. Specifically, ACC is calculated on D_u as illustrated in the following equation:

$$ACC = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbb{I}\left(\text{label}_i = OP\left(\hat{y}_i\right)\right). \tag{9}$$

In the equation, label_i represents the actual class label for $x_i \in D_u$, and this label is only provided during the testing phase. OP denotes the optimal permutation that aligns $\hat{y}_i$ with label_i. Our evaluation includes calculating the accuracy (ACC) for all samples (A), samples from known classes (K) in D_u, and samples from unknown classes (U) in D_u. This protocol is applied to all compared methods.

Implementation Details. We utilize the ViT-B/16 model [7], pre-trained with CLIP [36], as the backbone network for both image and text encoders and keep them frozen. To ensure the generalization of our method, unless otherwise specified, we adopt the following unified hyperparameter settings across all datasets: $\tau_T = 0.01$, $\tau_I = 0.05$, $k = 5$, $\beta = 0.5$. Textual Prompt Optimization and Bimodal Graph Enhancement modules are applied sequentially, with learnable parameters $Prompt = \{u_i\}_{i=1}^{|C_u|}$ and $\hat{Y}$, respectively. All experiments are conducted on a single NVIDIA A6000 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 different method. To ensure fairness, we use the ViT-B/16 model as the teacher model for the compared methods. For the TP-OWSSL method,

the memory required to run it on the ImageNet-1k and WebVision datasets exceeds the available GPU memory, so we do not include it in the comparison.

Experimental results indicate that our method outperforms state-of-the-art OWSSL methods across multiple evaluation metrics. This performance gain stems from several key design choices: (1) Textual prompt optimization mitigates model bias towards known classes during tuning, enabling the generation of less biased and more effective textual representations for unknown classes, a critical step often overlooked in existing methods. (2) Bimodal graph enhancement constructs a bi-modal proximity graph from image-image and image-prompt similarities, effectively mitigating the graph sparsity issue in prior work. (3) Label propagation algorithm iteratively updates the pseudo-label confidence matrix, effectively facilitating propagation of visual and semantic information across the graph. These strategies significantly enhance the model's ability to discover unknown classes.

4.3 Further Analyses

In this section, we provide empirical evidence to validate the contributions of these strategies.

Analysis of Effectiveness of Textual Prompt Optimization. Existing text-guided OWSSL methods often overlook model bias towards known classes during textual prompt tuning [8,9]. This bias inclines the model to classify samples as known classes, thereby suppressing the optimization process for learning effective textual representations of unknown classes. To address this, we introduce a textual prompt optimization strategy that mitigates this bias, enabling the optimized prompts to generate less biased and effective textual representations for unknown classes.

Table 3: Comparison of the count of unknown samples misclassified as known classes across different methods.

Datasets	Methods	Mis-Count
Hybrid	CSC-OWSSL	2.9k
	TPOBGE	1.5k
ImageNet-1K	CSC-OWSSL	15k
	TPOBGE	1.6k

To validate the effectiveness of our strategy, we first compare the overall performance of our method against CSC-OWSSL. As shown in Table 2, our method achieves a substantial improvement. To understand the mechanism behind this gain, we analyze the number of unknown samples misclassified as known classes. The results in Table 3 demonstrate that our strategy significantly reduces such misclassifications, directly confirming its success in mitigating model bias. Furthermore, an ablation study within our method in Figure 3, confirms the indispensable contribution of this strategy to the final performance.

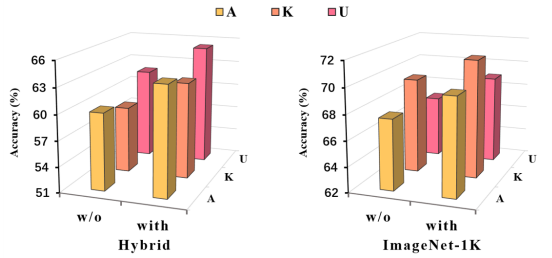

Fig. 3: Performance comparison between methods without and with textual prompt optimization.

Analysis of Effectiveness of Bimodal Graph Enhancement. Existing text-guided OWSSL methods often establish only limited cross-modal connections, resulting in a sparse graph that hinders the propagation of visual and semantic information. To address this, we employ bimodal graph enhancement, which involves constructing a proximity graph for image-image and image-prompt pairs to alleviate the graph sparsity issue. Additionally, label propagation algorithm is applied to facilitate the effective propagation of visual and semantic information across the graph.

Table 4: Comparison of the edge count across different methods.

Datasets	Methods	Edge Count
Hybrid	CSC-OWSSL	15.1k
	TPOBGE	114k
ImageNet-1K	CSC-OWSSL	50k
	TPOBGE	375k

To validate the effectiveness of this strategy, we first provide quantitative evidence from the graph structure itself. As shown in Table 4, applying bimodal graph enhancement leads to a significant increase in the number of edges. This stands in stark contrast to the sparse structure of CSC-OWSSL, where visual features are connected solely to their corresponding textual descriptions. Moreover, we further evaluate its performance contribution through ablation studies. As illustrated in Figure 4, the method incorporating bimodal graph enhancement outperforms the one using only textual prompt optimization.

Analysis of the Sensitivity of Hyperparameters. To ensure the good generalization ability of our method, we adopt a unified set of hyperparameters across all datasets: $\tau_T = 0.01$, $\tau_I = 0.05$, $k = 5$, $\beta = 0.5$. To verify the robustness of our method to different values of these hyperparameters, we adjust τ_T, τ_I, k, β individually while keeping all other parameters constant. We demonstrate the model's performance across different values of hyperparameters, and the rel-

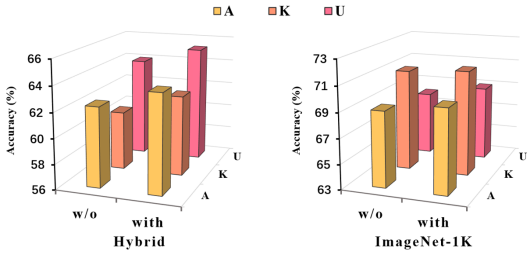

Fig. 4: Performance comparison between methods without and with bimodal graph enhancement.

evant experimental results are shown in Figure 5. These analyses demonstrate that our method is robust to different values of hyperparameters.

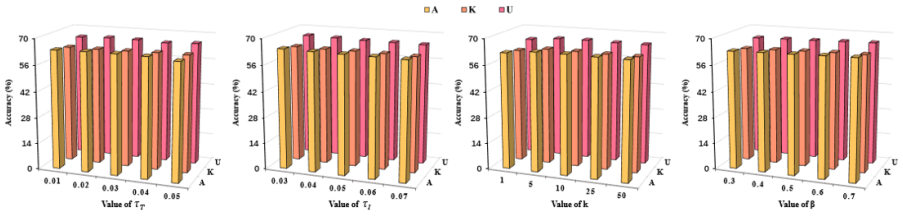

Fig. 5: Performance of TPOBGE with different values of hyperparameters on Hybrid dataset.

Analysis of Efficiency of Different OWSSL Methods. A key metric for applying OWSSL methods to practical tasks is efficiency and low resource consumption. Our method achieves this by freezing the parameters of the image and text encoders, requiring only a single pass through the image-text encoder for image and textual descriptions, and then optimizing only the prompts. We compare the training time and required GPU resources of CSC-OWSSL and our method across different datasets, choosing it as a benchmark for its leading performance and efficiency. The training time and resource consumption of CSC-OWSSL are on par with those of the lightweight baseline SimGCD. As shown in Table 5, our method requires significantly less training time and GPU resources than CSC-OWSSL. This significant improvement in efficiency makes our method more suitable for resource-constrained open environments.

5 Discussions

Limitations. Despite its effectiveness in the text-guided OWSSL setting, our method still exhibits limitations. First, the performance gains rely on sufficient

Table 5: Comparison of the training times and GPU resources of different methods.

Datasets	Methods	Training time	GPU resources
Hybrid	CSC-OWSSL	543 minutes	18970.10 MB
	TPOBGE	3 minutes	953.23 MB
ImageNet-1K	CSC-OWSSL	7303 minutes	18932.57 MB
	TPOBGE	10 minutes	9843.38 MB

training samples and the generalization capability of pre-trained VLMs, limiting its applicability in few-shot scenarios or vertical domains lacking strong prior knowledge. Second, the alignment of visual-textual modalities is only superficially addressed, and the deep semantic gap remains fundamentally unresolved.

Future Directions. This paper focuses on OWSSL, and the assumptions used represent the standard setting adopted for fair comparison, while real-world application scenarios are often more complex. Therefore, future research could proceed in the following directions: first, enhancing the method's robustness in more challenging realistic scenarios, such as those involving long-tailed distributions or out-of-distribution samples, to overcome the limitations of standard settings; and second, applying the method to more complex tasks, such as object detection and video understanding. These directions will advance OWSSL toward more robust and generalizable open-world intelligent systems.

6 Conclusion

In this paper, we propose a novel OWSSL method consisting of textual prompt optimization and bimodal graph enhancement. This effectively mitigates model bias towards known classes to generate less biased and more effective representations, and alleviates the sparsity issue in the graph structure to facilitate the effective propagation of visual and semantic information across the graph, thereby enhancing the model's ability to recognize classes. Experimental results on multiple datasets demonstrate that our method achieves higher performance than state-of-the-art OWSSL methods, particularly on datasets with a large number of classes. Future work will focus on enhancing the method's robustness in more challenging realistic scenarios, such as long-tailed distributions and the continuous emergence of unknown classes, and extending it to more complex downstream tasks like object detection and video understanding, thereby broadening its potential applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62541608, 62376141), the Shanxi Province Natural Science Foundation for Youths, China (No. 202503021212096), and the Fundamental Research Program of Shanxi Province (No. 202403021211086).

References

1. Cao, K., Brbic, M., Leskovec, J.: Open-world semi-supervised learning. In: Proceedings of the International Conference on Learning Representations (2022)
2. Cao, X., Zheng, X., Wang, G., Yu, W., Shen, Y., Li, K., Lu, Y., Tian, Y.: Solving the catastrophic forgetting problem in generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16880–16889 (2024)
3. Caselli, L., Mistretta, M., Magistri, S., Bagdanov, A.D.: Spectralgcd: Spectral concept selection and cross-modal representation learning for generalized category discovery. In: Proceedings of the International Conference on Learning Representations (2026)
4. Chi, H., Liu, F., Yang, W., Lan, L., Liu, T., Han, B., Niu, G., Zhou, M., Sugiyama, M.: Meta discovery: Learning to discover novel classes given very limited data. In: Proceedings of the International Conference on Learning Representations (2022)
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 2292–2300 (2013)
6. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (2021)
8. Fan, Y., Cui, J., Liang, J.: Learning textual prompts for open-world semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14756–14765 (2025)
9. Fan, Y., Cui, J., Liang, J., Liang, J.: Open-world semi-supervised learning with class semantic correlations. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 5092–5100 (2025)
10. Fang, C., Liang, J., Liang, J., Du, Z., Cao, F.: Point cloud semantic scene completion with prototype-guided transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3822–3830 (2026)
11. Fang, C., Liang, J., Liang, J., Wang, H., Yao, K., Cao, F.: Multi-modal point cloud completion with interleaved attention enhanced transformer. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 963–971 (2025)
12. Guo, L., Li, Y.: Class-imbalanced semi-supervised learning with adaptive thresholding. In: Proceedings of the International Conference on Machine Learning. pp. 8082–8094 (2022)
13. Guo, L., Zhang, Z., Jiang, Y., Li, Y., Zhou, Z.: Safe deep semi-supervised learning for unseen-class unlabeled data. In: Proceedings of the International Conference on Machine Learning. pp. 3897–3906 (2020)
14. Han, K., Li, Y., Vaze, S., Li, J., Jia, X.: What's in a name? beyond class indices for image recognition. CoRR **abs/2304.02364** (2023)
15. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8400–8408 (2019)
16. He, W., Meng, X., Huang, Z., Qi, X., Xiao, R., Li, C.: Multi-modal representation learning via semi-supervised rate reduction for generalized category discovery.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 39637–39646 (2026)
17. He, Z., Liu, Y., Han, K.: Category discovery: An open-world perspective. CoRR **abs/2509.22542** (2025)
18. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5070–5079 (2019)
19. Iscen, A., Tolias, G., Avrithis, Y., Furon, T., Chum, O.: Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 926–935 (2017)
20. Kim, H., Suh, S., Kim, D., Jeong, D., Cho, H., Kim, J.: Proxy anchor-based unsupervised learning for continuous generalized category discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16642–16651 (2023)
21. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 554–561 (2013)
22. Lafon, M., Ramzi, E., Rambour, C., Audebert, N., Thome, N.: Gallop: Learning global and local prompts for vision-language models. In: Proceedings of the European Conference on Computer Vision. pp. 264–282 (2024)
23. LeCun, Y., Bengio, Y., Hinton, G.E.: Deep learning. Nature **521**(7553), 436–444 (2015)
24. Li, W., Wang, L., Li, W., Agustsson, E., Gool, L.V.: Webvision database: Visual learning and understanding from web data. CoRR **abs/1708.02862** (2017)
25. Li, Y., Guo, L., Zhou, Z.: Towards safe weakly supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 334–346 (2021)
26. Li, Z., Meinel, C., Yang, H.: Generalized categories discovery for long-tailed recognition. CoRR **abs/2401.05352** (2024)
27. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 17612–17625 (2022)
28. Ma, S., Zhu, F., Zhang, X., Liu, C.: Protogcd: Unified and unbiased prototype learning for generalized category discovery. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(7), 6022–6038 (2025)
29. Ma, S., Zhu, F., Zhong, Z., Liu, W., Zhang, X., Liu, C.: Happy: A debiased learning framework for continual generalized category discovery. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 50850–50875 (2024)
30. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729 (2008)
31. Oliver, A., Odena, A., Raffel, C., Cubuk, E.D., Goodfellow, I.J.: Realistic evaluation of deep semi-supervised learning algorithms. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 3239–3250 (2018)
32. Ouldnoughi, R., Kuo, C., Kira, Z.: CLIP-GCD: simple language guided generalized category discovery. CoRR **abs/2305.10420** (2023)
33. Pu, N., Li, W., Ji, X., Qin, Y., Sebe, N., Zhong, Z.: Federated generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28741–28750 (2024)

34. Pu, N., Zhong, Z., Sebe, N.: Dynamic conceptional contrastive learning for generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7579–7588 (2023)
35. Qian, Q., Xu, Y., Hu, J.: Intra-modal proxy learning for zero-shot visual categorization with CLIP. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 25461–25474 (2023)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763 (2021)
37. Rastegar, S., Salehi, M., Asano, Y.M., Doughty, H., Snoek, C.G.M.: Selex: Self-expertise in fine-grained generalized category discovery. In: Proceedings of the European Conference on Computer Vision. pp. 440–458 (2024)
38. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(7), 1757–1772 (2013)
39. Shi, L., Zhen, H., Zhang, G., Yan, J.: Relative entropic optimal transport: a (prior-aware) matching perspective to (unbalanced) classification. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 22085–22098 (2023)
40. Shin, C., Zhao, J., Cromp, S., Vishwakarma, H., Sala, F.: OTTER: effortless label distribution adaptation of zero-shot models. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 44064–44101 (2024)
41. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 596–608 (2020)
42. Stojnic, V., Kalantidis, Y., Tolias, G.: Label propagation for zero-shot classification with vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23209–23218 (2024)
43. Troisemaine, C., Lemaire, V., Gosselin, S., Reiffers-Masson, A., Flocon-Cholet, J., Vaton, S.: Novel class discovery: an introduction and key concepts. CoRR **abs/2302.12028** (2023)
44. Udandarao, V., Gupta, A., Albanie, S.: Sus-x: Training-free name-only transfer of vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2725–2736 (2023)
45. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7482–7491 (2022)
46. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. California Institute of Technology (2011)
47. Wang, H., Vaze, S., Han, K.: Hilo: A learning framework for generalized category discovery robust to domain shifts. In: Proceedings of the International Conference on Learning Representations (2025)
48. Wen, X., Zhao, B., Qi, X.: Parametric classification for generalized category discovery: A baseline study. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16544–16554 (2023)
49. Yang, M., Yin, J., Gu, Y., Deng, C., Zhang, H., Zhu, H.: Consistent prompt tuning for generalized category discovery. International Journal of Computer Vision **133**(7), 4014–4041 (2025)

50. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 18408–18419 (2021)
51. Zheng, H., Pu, N., Li, W., Sebe, N., Zhong, Z.: Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In: Proceedings of the European Conference on Computer Vision. pp. 41–58 (2024)
52. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 321–328 (2003)
53. Zhu, F., Ma, S., Cheng, Z., Zhang, X., Zhang, Z., Liu, C.: Open-world machine learning: A review and new outlooks. CoRR **abs/2403.01759** (2024)