PromptCAL: Contrastive Affinity Learning via Auxiliary Prompts for Generalized Novel Category Discovery

Sheng Zhang[†✉] Salman Khan[†◇] Zhiqiang Shen[†⋆] Muzammal Naseer[†]
Guangyi Chen[†‡] Fahad Shahbaz Khan[†*]
[†]Mohamed bin Zayed University of AI [⋆]Hong Kong University of Science and Technology
[◇]Australian National University [*]Linköping University [‡]Carnegie Mellon University
{firstname.lastname}@mbzuai.ac.ae

Abstract

Although existing semi-supervised learning models achieve remarkable success in learning with unannotated in-distribution data, they mostly fail to learn on unlabeled data sampled from novel semantic classes due to their closed-set assumption. In this work, we target a pragmatic but under-explored Generalized Novel Category Discovery (GNCD) setting. The GNCD setting aims to categorize unlabeled training data coming from known and novel classes by leveraging the information of partially labeled known classes. We propose a two-stage Contrastive Affinity Learning method with auxiliary visual Prompts, dubbed PromptCAL*, to address this challenging problem. Our approach discovers reliable pairwise sample affinities to learn better semantic clustering of both known and novel classes for the class token and visual prompts. First, we propose a discriminative prompt regularization loss to reinforce semantic discriminativeness of prompt-adapted pre-trained vision transformer for refined affinity relationships. Besides, we propose contrastive affinity learning to calibrate semantic representations based on our iterative semi-supervised affinity graph generation method for semantically-enhanced supervision. Extensive experimental evaluation demonstrates that our PromptCAL method is more effective in discovering novel classes even with limited annotations and surpasses the current state-of-the-art on generic and fine-grained benchmarks* (e.g.*, with nearly* 11% *gain on CUB-200, and* 9% *on ImageNet-100) on overall accuracy. Our code is available at https://github.com/sheng-eatamath/PromptCAL.*

1. Introduction

The deep neural networks have demonstrated favorable performance in the *Semi-Supervised Learning* (SSL) setting [15,29,40,46,49]. Some recent works can even achieve comparable performance to their fully-supervised counterparts using few annotations for image recognition [3,38,46]. However, these approaches heavily rely on the *closed-world* assumption that unlabeled data share the same underlying class label space as the labeled data [10,47]. In many realistic scenarios, this assumption does not hold true because of the inherent dynamism of real-world tasks where novel classes can emerge in addition to known classes.

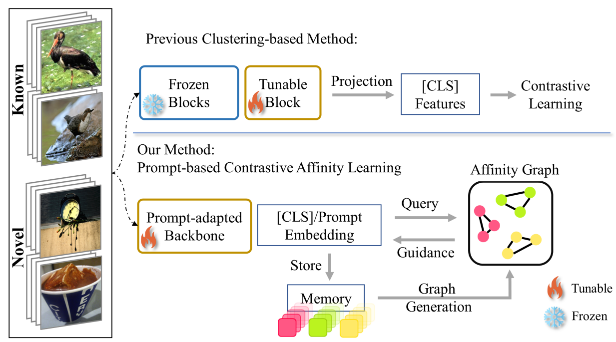

Figure 1. **PromptCAL Overview.** In contrast to previous method based on semi-supervised contrastive learning, PromptCAL constructs affinity graph on-the-fly to guide representation learning of the class token and prompts. Meanwhile, our prompt-adapted backbone can be tuned to enhance semantic discriminativeness. PromptCAL can discover reliable affinities from a memory bank, especially for novel classes. Therefore, our PromptCAL is better task-aligned and discriminative to novel semantic information.

In contrast to SSL, the *Novel Category Discovery* (NCD) problem was introduced by [12] to relax the closed-world assumption of SSL, which assumes the unlabeled data contain novel classes. Recently, the nascent *Generalized Novel Category Discovery* (GNCD) problem, first proposed in [4, 41], extends NCD and assumes the unlabeled data can contain both known and novel classes, which is more pragmatic and challenging. To be more specific, GNCD intends to categorize images sampled from predefined categories in the training set comprising labeled-knowns, unlabeled-knowns, and unlabeled-novels.

Our work focuses on GNCD problem. The key challenge of GNCD is to discriminate among novel classes when only the ground truths of known classes are accessible in the training set. Recent studies show that self-supervised pre-trained representations are conducive to discovering novel semantics [4, 5, 11, 41, 53]. A typical work on GNCD [41] takes advantage of the large-scale pre-trained visual transformer (ViT) [37], and learns robust clusters for known and novel classes through semi-supervised contrastive learning on downstream datasets. However, we discover that the remarkable potential of pre-trained ViT is actually suppressed by this practice, due to the *class collision* [52] issue induced by abundant false negatives in contrastive loss, *i.e.*, considering different unlabeled images from the same or similar semantic class as false negatives. As supported by empirical studies, abundant false negatives in contrastive training can deteriorate the compactness and purity of semantic clustering [5, 16, 21, 52]. Based on empirical investigation, we show that this issue is particularly severe in category discovery. Furthermore, although the existing commonly adopted practice [4, 41] of freezing most parts of the pre-trained backbone can alleviate overfitting on known classes, it constrains the flexibility and adaptability of backbones [18]. Lack of adaptability inhibits models from learning discriminative semantic information on downstream datasets.

To address above limitations and learn better semantically discriminative representations, we propose **Prompt**-based **C**ontrastive **A**ffinity **L**earning (**PromptCAL**) framework to tackle GNCD problem. To be specific, our approach aims to discover semantic clusters in unlabeled data by simultaneous semantic prompt learning based on our Discriminative Prompt Regularization (DPR) loss and representation calibration based on our Contrastive Affinity Learning (CAL) process. *Firstly*, CAL discovers abundant reliable pseudo positives for DPR loss and contrastive loss based on generated affinity graphs. These semantic-aware pseudo labels further enhance the semantic discriminativeness of DPR supervision. *Secondly*, DPR regularizes semantic representations of ensembled prompts, which facilitates the discovery of more accurate pseudo labels at the next-step of CAL. Therefore, as model and prompt representations are iteratively enhanced, we can obtain higher quality pseudo positives for further self-training as well as acquire better semantic clustering.

Our PromptCAL achieves State-Of-The-Art (SOTA) performance in extensive experimental evaluation on six benchmarks. Specifically, PromptCAL remarkably surpasses previous SOTA by more than 10% clustering accuracy on the fine-grained CUB-200 and StandfordCars datasets; it also significantly outperforms previous SoTAs by nearly 4% on ImageNet-100 and 7% on CIFAR-100. Interestingly, we identify that both DPR supervised prompts and unsupervised prompts of PromptCAL can learn semantic discriminativeness, which advances the flexibility of the pre-trained backbone. Furthermore, PromptCAL still achieves the best performance in challenging low-labeling and few-class setups.

Our contributions include: **(1)** We propose a two-stage framework for the generalized novel category discovery problem, in which semantic prompt tuning and contrastive affinity learning mutually reinforce and benefit each other during the learning process. **(2)** We propose two synergistic learning objectives, contrastive affinity loss and discriminative prompt regularization loss, based on our semi-supervised adapted affinity graphs to enhance semantic discriminativeness. **(3)** We comprehensively evaluate our method on three generic (*i.e.*, CIFAR-10, CIFAR-100, and ImageNet-100) and three fine-grained benchmarks (*i.e.*, CUB-200, Aircraft, and StandfordCars), achieving state-of-the-art performance, thereby showing its effectiveness. **(4)** We further showcase generalization ability of PromptCAL and its effectiveness in more challenging low-labeling and few-class setups.

2. Related Work

Category discovery. *Novel Category Discovery* (NCD), first formulated by DTC [12], aims to categorize the unlabeled novel classes by transferring the knowledge from labeled known classes [9, 11, 12, 45, 51, 53, 54]. The challenging NCD differs from SSL [40] in that the unlabeled data are sampled from distinct underlying semantic distribution. DTC [12] proposes to jointly warm up network weights and cluster prototypes based on DEC [45] method on unlabeled data, and then fit an annealing sharpened distribution. RankStats [11] and RS+ [51] propose to utilize ranking statistics to generate pseudo positives among unlabeled novel classes. OpenMix [54] transfers semantic knowledge by MixUp augmentation [50] between known and novel classes as well as between reliable novel anchors and other novel examples. NCL [53] proposes a neighborhood contrastive loss and a hard-negative generation process by mixing [50] novel and known classes. UNO [9] first formulates the NCD problem into classification based on dynamic class assignments by Sinkhorn-Knopp algorithm [23]. WTA [19] addresses multi-modal novel category discovery by inter- and intra-modal contrastive learning with permutation-ensembled ranking statistics as the pseudo-labeling method.

Generalized Novel Category Discovery (GNCD) problem, first proposed in [41], further extends NCD under a more realistic assumption that unlabeled data can be both sampled from novel classes and known classes. Specifically, the model learns to categorize unlabeled training data containing known and novel classes based on the knowledge of labeled known classes. Besides, a concurrent work, ORCA [4] proposes an uncertainty adaptive margin loss to reduce the intra-class variances between known and novel classes. GCD [41] addresses this challenging problem via proposed semi-supervised contrastive learning on large-scale pre-trained visual transformer (ViT) followed by constraint KMeans [1, 2]. However, GCD still has limitations: first, the frozen backbone lacks the adaptability to downstream tasks; besides, abundant false negatives will degenerate the semantic representation [7, 16, 21, 52]. To fully unleash the potential of pre-trained ViT, we address these two critical issues via our proposed prompt-based contrastive affinity learning.

Positive Mining in Neighborhoods. Some recent works in self-supervised learning discovered that mining positives to antagonize the side effect of abundant false negatives in the sample-wise contrastive loss is essential to the downstream performance [8, 16, 21, 35, 52]. FNC [16] comprehensively analyzes the adverse effect of false-negatives on contrastive learning SoTAs and performs positive mining based on ensembled similarities on local patch pairs. LA [55] proposes to learn better representation through soft clusters in neighborhoods at different scales. NNCLR [8], NCL [53], and WCL [52] conduct positive mining based on K-Nearest Neighbors (KNN) as pseudo positives to improve contrastive learning. We find one work in SSL [17] also leverages a graph diffusion algorithm to propagate pseudo labels. But there exist major differences between their work and ours: first, features in our context are prone to open-set noises [10] and thus more challenging than SSL; second, we conduct an efficient online diffusion per iteration via a graph subsampling strategy, while they conduct diffusion per epoch on the entire dataset; third, we compute affinity propagation on consensus affinity graph with prior knowledge, while they conduct propagation on naive KNN graph. Our framework incorporates and generalizes consensus KNN [34], which was originally built upon non-learnable SIFT [31] features of synthetic datasets, while our method exploits deep features and can guide end-to-end training, which suits the GNCD context.

Visual prompt learning. Prompt learning originates from the field of Natural Language Processing (NLP) [30]. Visual prompt learning (VPT) [18] tunes embedded visual prompts with a frozen pre-trained ViT backbone supervised by downstream objectives, which achieves better transfer. However, based on our experimental analysis, VPT [18] does not exhibit significant benefits especially on fine-grained datasets. Our objective acts as prompt regularization or a weaker semantic supervision signal, which is distinct from the learning goals of prompt ensembling [20, 36] and prompt composition [13] in NLP [30].

3. Method

The challenging aspect of GNCD in comparison to SSL is clustering novel semantics under both semantic shifts and missing annotations [40, 47]. However, existing methods [11, 41, 51, 53] cannot reliably discover and employ semantic affinities on pre-trained representations. Meanwhile, recent SoTAs [4, 41] lack suitable strategies to adapt the pre-trained backbone to learn discriminative semantic information without overfitting on known classes.

To this end, we propose PromptCAL, which consists of two synergistic learning objectives: discriminative prompt regularization (DPR) and contrastive affinity learning (CAL). The whole framework is displayed in Fig. 2. Specifically, in the first stage, we learn warm-up representation (in Sec. 3.2) for further tuning. Our DPR loss which is applied to both stages for prompt regularization is also explained. In the second stage, we discover reliable pseudo positives on generated affinity embedding graphs based on semi-supervised affinity generation (SemiAG) mechanism (in Sec. 3.3). Next, we propose our contrastive affinity loss (in Sec. 3.4) on pseudo labels generated by online SemiAG with the support of embedding memories. Lastly, we also present PromptCAL training algorithm in Appendix E.

3.1. Preliminaries

Before introducing our method, we formulate the GNCD problem and present some preliminaries.

Problem Definition. Our GNCD setting follows [41]. Specifically, we assume that the training dataset $\mathcal{D} = \mathcal{D}_l \bigcup \mathcal{D}_u$ comprises two subsets: a labeled set $\mathcal{D}_l = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N_1} \subset \mathcal{X}_l \times \mathcal{Y}_l$ with its label space $\mathcal{Y}_l = \mathcal{C}_{kwn}$, and an unlabeled set $\mathcal{D}_u = \{\boldsymbol{x}_i\}_{i=1}^{N_2} \subset \mathcal{X}_u$ with its underlying label space $\mathcal{Y}_u = \mathcal{C} = \mathcal{C}_{kwn} \bigcup \mathcal{C}_{new}$. Here, $\mathcal{C}$, $\mathcal{C}_{kwn}$, and $\mathcal{C}_{new}$ denote the label set for All, Known, and New classes, respectively. Following [41], we assume $|\mathcal{C}|$ is known.

Architecture. We take a self-supervised pre-trained ViT as our backbone [37]. We denote our visual prompt-adapted ViT backbone [18] as $f(\cdot|\theta, \theta_{\text{P}})$ parameterized by prompts θ_{P} and last block weights θ. In each mini-batch $\mathcal{B}$, there are two augmented views for each sample. Given a sample vector $\mathbf{x} \in \mathcal{B}$, we can extract its embedding $\mathbf{h} = f(\mathbf{x}|\theta, \theta_{\text{P}}) \in \mathcal{H}$ and project $\mathbf{h}$ into feature vector $\mathbf{z} = g(\mathbf{h}|\theta_{\text{H}}) \in \mathcal{Z}$ through a projection head $g(\cdot|\theta_{\text{H}})$ with parameters θ_{H}. Here, $\mathcal{H}, \mathcal{Z}$ denote embedding and feature spaces.

Contrastive Loss. To simplify notations of PromptCAL, we extend the definition of the standard supervised contrastive loss [22] as follows. Given a l_2-normalized query vector $\mathbf{t}_q$ and a set of l_2-noramlized key vectors $\mathbf{T}_k$ (which can be from the embedding or feature space), we define:

$$L_{\text{con}}(\mathbf{t}_q, \mathbf{T}_k; \tau, \mathcal{P}, \mathcal{A}) = -\frac{1}{|\mathcal{P}(\mathbf{t}_q)|} \sum_{\mathbf{t}_k^+ \in \mathcal{P}(\mathbf{t}_q)} \frac{\exp(\frac{\mathbf{t}_q \cdot \mathbf{t}_k^+}{\tau})}{\sum_{\mathbf{t_a} \in \mathcal{A}(\mathbf{t}_q)} \exp(\frac{\mathbf{t}_q \cdot \mathbf{t_a}}{\tau})} \quad (1)$$

where τ is the temperature parameter of the contrastive loss, and $\cdot$ denotes the cosine similarity operation. Here, $\mathcal{P}(\mathbf{t}_q)$

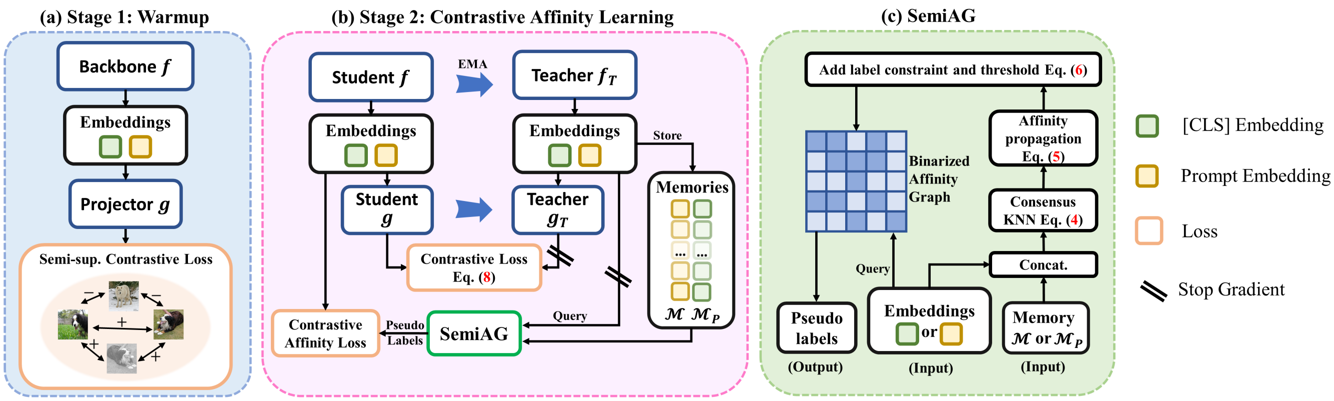

Figure 2. **Overview of our PromptCAL framework.** Our prompt-adapted backbone outputs a class embedding and ensembled prompt embedding. (a) In warm-up training, we conduct semi-supervised contrastive clustering (Semi-sup. Contrastive Loss) on the projected features of the class token and ensembled prompt, respectively. (b) In contrastive affinity learning stage, at each iteration, we forward the student and EMA (exponentially moving averaged) teacher with different augmented views of images. Output teacher embeddings are enqueued into their corresponding token-specific memory. We iteratively compute semi-supervised contrastive loss on the current batch and our contrastive affinity loss for student embeddings and memory embeddings with pseudo-labels from the dynamically generated affinity graph by SemiAG. (c) We generate affinity graphs for the class embedding and prompt embedding respectively via affinity propagation with label constraints on their corresponding consensus KNN graphs.

and $\mathcal{A}(\mathbf{t}_q)$ represent the positive set and anchor set of the query $\mathbf{t}_q$, respectively, which are subsets of $\mathbf{T}_k$.

3.2. Warm-up Phase with Discriminative Prompt Regularization

Discriminative Prompt Regularization. Although computation overheads are largely reduced by only tuning the last block [41], it restricts the backbone from better learning semantic representations and adapting to diverse downstream datasets. Counterintuitively, we discover that naively adapting backbone with visual prompts [18] overfits small datasets (refer to ablations on CUB-200 [42] in Sec. 4.5).

Motivated by [27, 43], we propose a discriminative prompt regularization loss to regularize and force prompts to learn semantically discriminative features with a task-related auxiliary loss. We investigate the superiority of DPR supervision on our prompt-adapted backbone in ablation study (Sec. 4.5) and Appendix D.

We assign input prompts at the last ViT block as [P] tokens (short for prompt), the output of which are ensembled and supervised by a task-related clustering loss in both training stages. All the remaining prompts are unsupervisedly learned, which provides the backbone with extra flexibility. Concretely, we average the l_2-normalized output embeddings of all [P] tokens into an ensembled embedding $\mathbf{h}_{\text{P}}$ (the same shape as the class embedding), and forward it to the projection head and obtain $\mathbf{z}_{\text{P}}$. Finally, we define the DPR task-related loss function on $\mathbf{h}_{\text{P}}/\mathbf{z}_{\text{P}}$ as the same form of the loss defined on $\mathbf{h}/\mathbf{z}$ but with a weaker weight γ.

Warm-up Training. Since randomly initialized prompts are not ready for contrastive affinity learning, we include warm-up training to prepare the class token and prompts with dataset-specific representation. The overall training objective in this stage is formulated as:

$$L_1(\mathbf{x}) = L^{\text{CLS}}_{\text{semi}}(\mathbf{z}) + \gamma L^{\text{P}}_{\text{semi}}(\mathbf{z}_{\text{P}}) \tag{2}$$

where $L^{\text{CLS}}_{\text{semi}}$ and $L^{\text{P}}_{\text{semi}}$ represent the semi-supervised contrastive loss (SemiCL) on [CLS] and its DPR counterpart on [P], respectively. Here, γ is DPR loss weight. Further, based on extended contrastive loss (Eq. 1), the SemiCL on [CLS] feature $\mathbf{z} \in \mathbf{Z}_{\mathcal{B}}$ is written as:

$$\begin{aligned} L^{\text{CLS}}_{\text{semi}}(\mathbf{z}) = {} & (1-\alpha) L_{\text{con}}\Big(\mathbf{z}, \mathbf{Z}_{\mathcal{B}}; \tau, \mathcal{P}_{\text{self}}, \mathcal{A}_{\text{self}}\Big) \\ & + \alpha L_{\text{con}}\Big(\mathbf{z}, \mathbf{Z}_{\mathcal{B}_l}; \tau_a, \mathcal{P}_{\text{sup}}, \mathcal{A}_{\text{sup}}\Big) \mathbb{I}\Big(\mathbf{z} \in \mathbf{Z}_{\mathcal{B}_l}\Big) \end{aligned} \tag{3}$$

where τ, τ_a are temperature parameters, and $\mathbb{I}$ is an indicator function. The first and second terms denote self-supervised and supervised contrastive loss on projected features of an entire batch $\mathbf{Z}_{\mathcal{B}}$ and only labeled samples $\mathbf{Z}_{\mathcal{B}_l}$, respectively. Following [14, 22], we define $\mathcal{P}_{\text{self}}(\mathbf{z})$ as the augmented counterpart of $\mathbf{z}$ in $\mathbf{Z}_{\mathcal{B}}$, and define $\mathcal{P}_{\text{sup}}(\mathbf{z})$ as all other features in $\mathbf{Z}_{\mathcal{B}_l}$ that shares the same class label with $\mathbf{z}$. Besides, we have $\mathcal{A}_{\text{sup}}(\mathbf{z}) = \mathbf{Z}_{\mathcal{B}_l} - \{\mathbf{z}\}$ and $\mathcal{A}_{\text{self}}(\mathbf{z}) = \mathbf{Z}_{\mathcal{B}} - \{\mathbf{z}\}$. Similar to Eq. 3, we can define the DPR loss function $L^{\text{P}}_{\text{semi}}$ on ensembled prompt feature $\mathbf{z}_{\text{P}}$ in the overall loss (Eq. 2).

3.3. Semi-supervised Affinity Generation

Once the warmed-up semantic representation for the class token and prompts are obtained, abundant positive samples can be discovered by reliable pseudo-labeling methods for enhanced clustering and supervision signals at

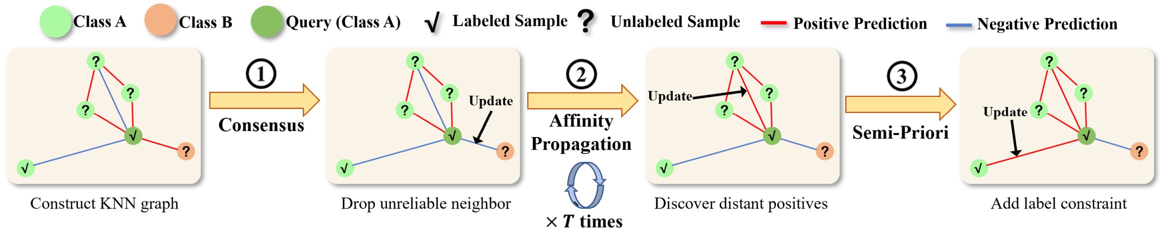

Figure 3. **An intuitive toy example for SemiAG**, which sequentially requires three operations. The relative pairwise distances are proportional to cosine distances in the embedding space. Each of the four graphs denotes results obtained at each step after binarization with thresholds. Each operation can either remove false positives or retrieve ground-truth positives for the query embedding (dark green). Firstly, only reliable neighbors are retrieved as positives based on consensus information; secondly, more positives are retrieved by affinity propagation on the entire graph; and thirdly, pairwise constraints in label information of labeled data (SemiPriori) are incorporated for affinity calibration.

next iteration. However, pseudo-labeling techniques in recent works (*e.g.*, naive nearest neighbors, pair-wise predictions as positives [4, 9, 11, 21, 53]) are not robust enough to semantic shifts [33]. To address this issue, we propose a semi-supervised affinity generation method under the assumption that consensus local neighbors share the same semantics. Specifically, we *first* construct an consensus affinity graph in $\mathcal{H}$ based on neighborhood statistics [34]. *Then*, we conduct affinity propagation on the entire graph to calibrate affinities. *Lastly*, we incorporate the semi-supervised priori from $\mathcal{D}_l$ into the graph. We explain these steps below. An illustrative example is presented in Fig. 3. The workflow of SemiAG operations is presented in Fig. 2 (c).

Consensus KNN graph. Given an embedding graph $\mathbf{G}_{\mathcal{H}} = (\mathcal{V}, \mathcal{E})$ whose node set $\mathcal{V} = \{\mathbf{h}_i\}_{i=1}^{N_G}$ contains N_G embeddings and edge set is $\mathcal{E} = \{e_{i,j} = \mathbf{h}_i \cdot \mathbf{h}_j\}_{i,j=1}^{N_G}$, we build a consensus graph $\mathbf{G}_c = (g_{i,j})_{i,j=1}^{N_G}$ on $\mathcal{V}$ via consensus statistics. Each edge $g_{i,j}$ of $\mathbf{G}_c$ is defined as:

$$g_{i,j} = \begin{cases} |\{\mathbf{h}_c | \mathbf{h}_i, \mathbf{h}_j \in \mathcal{O}_K(\mathbf{h}_c)), \forall \mathbf{h}_c \in \mathcal{V}\}| & i \neq j \\ 0 & i = j, \end{cases} \quad (4)$$

where $\mathcal{O}_K(\mathbf{h}_c) = \mathtt{argtopK}_{\mathbf{h}_j}(\{\mathbf{h}_j \cdot \mathbf{h}_c | \mathbf{h}_j \in \mathcal{V}\})$ denotes the K-neighborhood of $\mathbf{h}_c \in \mathcal{V}$. Then, we convert it into $\tilde{\mathbf{G}}_c$ by row normalization. However, consensus graph has a defect: the neighborhood consensus condition is rigorous and only considers local information, which means abundant potential positives are still unretrieved.

Affinity propagation with SemiPriori. To overcome this issue, we leverage the graph diffusion algorithm [48] on the probabilistic matrix $\tilde{\mathbf{G}}_c$ to propagate local affinities along multi-hop paths to characterize higher-order structural information and avoid degenerated solutions. Specifically, we apply TPG diffusion algorithm [48], which iteratively computes the diffused graph $\tilde{\mathbf{G}}_d$ as:

$$\tilde{\mathbf{G}}_d^{(t+1)} = \tilde{\mathbf{G}}_c \tilde{\mathbf{G}}_d^{(t)} \tilde{\mathbf{G}}_c^T + \mathbf{I}, t = 1, ..., \eta \quad (5)$$

where $\mathbf{I}$ is an identity matrix, and η is the total diffusion step. $\tilde{\mathbf{G}}_d^{(t)}$ denotes the t-th step diffused graph and $\tilde{\mathbf{G}}_d^{(0)} = \tilde{\mathbf{G}}_c$. We denote the final diffused graph as $\tilde{\mathbf{G}}_d$. In Appendix A, we provide more detailed descriptions.

However, the consensus graph and affinity propagation neglect abundant prior information in the labeled data. To address the issue, we incorporate SemiPriori, *i.e.*, add sample-wise class labels as pairwise constraints to $\tilde{\mathbf{G}}_d$. We set the edge to 1 if two nodes have the same labels (*i.e.*, $y_i = y_j$) and prune the edge if $y_i \neq y_j$. Meanwhile, we sparsify $\tilde{\mathbf{G}}_d$ with a pre-defined quantile q, then the generated binarized affinity graph $\mathbf{G}_b$ is denoted as:

$$\mathbf{G}_b(i,j) = \begin{cases} 1 & (y_i = y_j) \vee \left(\tilde{\mathbf{G}}_d(i,j) > q\right) \\ 0 & (y_i \neq y_j) \end{cases} \quad (6)$$

On binarized affinity graph $\mathbf{G}_b$, positive/negative pairs are regarded as reliable pseudo positives/negatives in noisy embedding space for further contrastive affinity learning (in Sec. 3.4). Therefore, pseudo-labels of both labeled and unlabeled data are computed; while, those of labeled data are calibrated by SemiPriori. Note that we compute two binarized graphs for [CLS] and [P] embeddings, respectively.

3.4. Contrastive Affinity Learning Phase

In this section, given reliable pseudo positives identified from an embedding graph, we introduce two critical components for the second phase learning: online graph sampling strategy and our proposed CAL loss. The overall framework of contrastive affinity learning is illustrated in Fig. 2 (b).

Graph sampling with memory. One practical issue arises (in Sec. 3.3): SemiAG on mini-batches is not effective due to sampling insufficiency; while conducting SemiAG offline on the entire dataset is time-consuming and memory inefficiency [17]. To strike a balance between the graph size and computation resources, inspired by [28], we dynamically construct a sub-graph $\mathbf{G}'_{\mathcal{H}}$ sub-sampled from the entire graph $\mathbf{G}_{\mathcal{H}}$ supported by an extra embedding memory bank $\mathcal{M}$ and an exponentially moving averaged (EMA) teacher ($f_{\mathrm{T}}, g_{\mathrm{T}}$), like MoCo [14]. Specifically, for each input batch, the EMA teacher produces stable embeddings, which are enqueued to the fixed-size first-in-first-out memory. The sub-graph $\mathbf{G}'_{\mathcal{H}}$ is then constructed by the embeddings in the memory and teacher embeddings in the current batch. We denote its node set as $\mathcal{V}(\mathbf{G}'_{\mathcal{H}}) = \mathcal{M} \bigcup \{\mathbf{h}_{\mathrm{T}} = f_{\mathrm{T}}(\mathbf{x}) | \mathbf{x} \in \mathcal{B}\}$. In this way, we can apply the same SemiAG operation (in Sec. 3.3) to the sub-graph on the fly with adjustable memory sizes. Note that we maintain another memory for SemiAG on prompts, since we retain DPR loss in contrastive affinity learning phase.

Contrastive affinity loss. The target of CAL loss is to gradually calibrate the semantic representation by learning from generated affinity constraints in graphs. Given the sub-graph $\mathbf{G}'_{\mathcal{H}}$ and its corresponding binarized graph $\mathbf{G}'_b$ by SemiAG (in Sec. 3.3), we formulate CAL loss with [CLS] embedding $\mathbf{h}_i$ as a query, embeddings in sub-graph node set $\mathcal{V}(\mathbf{G}'_{\mathcal{H}})$ as keys:

$$L^{\text{CLS}}_{\text{CAL}}(\mathbf{h}_i, \mathbf{G}'_b) = L_{\text{con}}(\mathbf{h}_i, \mathcal{V}(\mathbf{G}'_{\mathcal{H}}), \tau_a, \mathcal{P}_a, \mathcal{A}_a) \quad (7)$$

where τ_a is a hyper-parameter, and the positive set is defined as $\mathcal{P}_a(\mathbf{h}_i) = \{\mathbf{h}_{\mathrm{T},j} | \mathbf{G}'_b(i,j) = 1, \forall \mathbf{h}_{\mathrm{T},j\neq i} \in \mathcal{V}(\mathbf{G}'_{\mathcal{H}})\} \cup \{\mathbf{h}'_{\mathrm{T},i}\}$ where $\mathbf{h}'_{\mathrm{T},i}$ is $\mathbf{h}_i$ augmented counterpart. Note that $\mathcal{P}_a$ is always non-empty. Since the whole $\mathcal{V}(\mathbf{G}'_b)$ is too large, we define the anchor set $\mathcal{A}_a(\mathbf{h}_i)$ as the union of $\mathcal{P}_a(\mathbf{h}_i)$ and N_{neg} randomly sampled pseudo-negatives for each query. For $L^{\text{CLS}}_{\text{CAL}}$ loss, we also define its corresponding DPR counterpart of CAL loss as $L^{\text{P}}_{\text{CAL}}$.

Overall optimization objective. At CAL stage, we also preserve SemiCL loss in feature space to retain the model capability of instance-wise discrimination. To further increase the consistency between the teacher and student, we adapt supervised and self-supervised term of SemiCL (Eq. 3) as:

$$\begin{aligned} L^{\text{CLS}}_{\text{self}}(\mathbf{z}) &= L_{\text{con}}\Big(\mathbf{z}, \mathbf{Z}_{\mathcal{B},T}; \tau, \mathcal{P}_{\text{self}}, \mathcal{A}_{\text{self}}\Big) \\ L^{\text{CLS}}_{\text{sup}}(\mathbf{z}) &= L_{\text{con}}\Big(\mathbf{z}, \mathbf{Z}_{\mathcal{B}_l,T}; \tau_a, \mathcal{P}_{\text{sup}}, \mathcal{A}_{\text{sup}}\Big) \mathbb{I}\Big(\mathbf{z} \in \mathbf{Z}_{\mathcal{B}_l}\Big) \end{aligned} \quad (8)$$

Here, we use student feature $\mathbf{z}$ as a query and teacher features $\mathbf{Z}_{\mathcal{B},T}, \mathbf{Z}_{\mathcal{B}_l,T}$ as keys to strengthen consistencies. The positive and anchor sets follow the same definition as in Eq. (3) but are defined in the teacher feature space.

Then, the overall loss for [CLS] token at CAL stage is formulated as:

$$L^{\text{CLS}}_2 = (1-\alpha) L^{\text{CLS}}_{\text{sup}} + \alpha\Big(\beta L^{\text{CLS}}_{\text{CAL}} + (1-\beta) L^{\text{CLS}}_{\text{self}}\Big) \quad (9)$$

where β is an adjustable weight. Its corresponding DPR counterpart can be similarly defined, denoted as L^{P}_2.

Finally, since we also adopt DPR at CAL stage, the overall optimization objective is formulated as:

$$L_2 = L^{\text{CLS}}_2 + \gamma L^{\text{P}}_2 \quad (10)$$

During the inference, the [CLS] embeddings are adopted as final predictions.

4. Experiments

4.1. Datasets

We evaluate PromptCAL on three generic datasets (*i.e.*, CIFAR-10/100 [25] and ImageNet-100 [26]) and three fine-grained datasets (*i.e.*, CUB-200 [42], StandfordCars [24], and Aircraft [32]). A summary of datasets is listed in Appendix B. For each dataset, we first subsample $|\mathcal{C}_{kwn}|$ known classes from all classes. Then, a pre-defined ratio of images for known classes are sampled to form the labeled set $\mathcal{D}_l$. Follow GCD [41], we set labeling ratio to 80% for CIFAR-100 and 50% for other datasets unless otherwise specified. All unsampled images constitute $\mathcal{D}_u$. In practice, we adopt the same dataset split of $\mathcal{D}_l$ and $\mathcal{D}_u$ as in [41]. (See Table 6 in Appendix B for more details on known class numbers and labeling ratios for all dataset). Besides, we adopt fewer $|\mathcal{C}_{kwn}|$ and smaller labeling ratios in more challenging setups for ablation study (Sec. 5).

4.2. Evaluation Protocol

We follow GCD [41] evaluation protocol in all experiments unless otherwise specified. Specifically, we perform SemiKMeans clustering [41] on the predicted embeddings. Then, all clusters are mapped through the optimal assignment solved by Hungarian algorithm [44] to their ground-truth classes. The accuracy scores for All, Known, and New classes are reported. The predicted embeddings from the student class token are evaluated during inference.

4.3. Implementation Details

Following GCD [41], we use ViT-B/16 pre-trained DINO [6] on ImageNet-1K [26] as our backbone for evaluation. For all experiments, we fix the batch size to 128 and use the same data augmentation strategies as [41]. We present complete implementation details in Appendix C.

4.4. Main Results

Evaluation on generic datasets. We evaluate both stages of PromptCAL on three generic datasets (*i.e.*, CIFAR-10/100 [25], and ImageNet-100 [26]). Table 1 shows that our PromptCAL consistently and significantly surpasses previous SoTAs, *i.e.*, ViT-adapted ORCA [4], our baseline GCD [41], and adapted NCD SOTA methods (UNO+ [9] and RankStats+ [11]) in terms of overall accuracy on all three datasets. Specifically, PromptCAL surpasses GCD by 6.4% on CIFAR-10, 8.2% on CIFAR-100, and 9.0% ImageNet-100 on All classes; it also remarkably outperforms ORCA by 7% on CIFAR-100 and 3.9% on ImageNet-100. Besides, in contrast to ORCA and UNO+ which suffer from severe overfitting on Known classes, PromptCAL manifests substantial advantages over other methods on New classes (about 10% improvements on three datasets).

Method	CIFAR-10 All	CIFAR-10 Known	CIFAR-10 New	CIFAR-100 All	CIFAR-100 Known	CIFAR-100 New	ImageNet-100 All	ImageNet-100 Known	ImageNet-100 New
KMeans [1]	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ [11]	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [9]	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	**95.0**	57.9
GCD [41]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
ORCA† [4]	96.9	95.1	97.8	74.2	82.1	67.2	79.2	93.2	72.1
PromptCAL-1^{st} (Ours)	97.1	97.7	96.7	76.0	80.8	66.6	75.4	94.2	66.0
PromptCAL-2^{nd} (Ours)	**97.9**	96.6	**98.5**	**81.2**	**84.2**	**75.3**	**83.1**	92.7	**78.3**

Table 1. **Evaluation on three generic datasets.** Accuracy scores are reported. †denotes adapted methods. Both stages of PromptCAL are evaluated.

Method	CUB-200 All	CUB-200 Known	CUB-200 New	StanfordCars All	StanfordCars Known	StanfordCars New	Aircraft All	Aircraft Known	Aircraft New
KMeans [1]	34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8
RankStats+ [11]	33.3	51.6	24.2	28.3	61.8	12.1	27.9	**55.8**	12.8
UNO+ [9]	35.1	49.0	28.1	35.5	**70.5**	18.6	28.3	53.7	14.7
GCD [41]	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9
ORCA† [4]	36.3	43.8	32.6	31.9	42.2	26.9	31.6	32.0	31.4
PromptCAL-1^{st} (Ours)	51.1	55.4	48.9	42.6	62.8	32.9	44.5	44.6	44.5
PromptCAL-2^{nd} (Ours)	**62.9**	**64.4**	**62.1**	**50.2**	70.1	**40.6**	**52.2**	52.2	**52.3**

Table 2. **Evaluation on three fine-grained datasets.** Accuracy scores are reported. †denotes adapted methods. Both stages of PromptCAL are evaluated.

By comparing the 1^{st} stage (PromptCAL-1^{st}) with the 2^{nd} stage (PromptCAL-2^{nd}), we observe major performance boosts, especially on `New` classes. In addition, we also notice that both stages of our PromptCAL have significant contributions to the final performance on generic datasets. Specifically, PromptCAL-1^{st} improves 5.6% and 3.0% over GCD on CIFAR-10/100, respectively; while the PromptCAL-2^{nd} further improves by 5.2% and 9.0% on CIFAR-100 and ImageNet-100, respectively. Besides, we also achieve $\sim$ 7% boost of overall accuracy on CIFAR-100 and 4% on ImageNet-100 when compared with ORCA. Therefore, above results validate advantages and effectiveness of our two-stage PromptCAL in category discovery.

Evaluation on fine-grained datasets. We also report results on fine-grained datasets to demonstrate the PromptCAL effectiveness in Table 2. Apparently, the low performance of KMeans illustrates the challenging nature of fine-grained category discovery caused by larger intra-class and lower inter-class variations. Notice that ORCA performance degrades substantially on three fine-grained datasets. In contrast, our PromptCAL consistently exceeds NCD SOTA and ORCA, and outperforms GCD by $\sim$ 11% on `All` classes on CUB-200 and StanfordCars and $\sim$ 7% on Aircraft. Different from results in Table 1, the results on fine-grained datasets show that the major performance gain of PromptCAL originates from the 2^{nd} CAL stage. Noticeably, PromptCAL-1^{st} performance even drops compared with GCD on CUB-200 and Aircraft datasets; while, PromptCAL-2^{nd} achieves remarkable and consistent improvements, especially on `New` classes.

4.5. Ablation and analysis

In this section, we conduct extensive ablation experiments to reveal and investigate contributions of each component. Next, we present in-depth analysis on the effectiveness of SemiAG and discuss the effect of visual prompts in PromptCAL. Further, we explore how PromptCAL performs in more challenging and real-world scenarios with lower-labeling and fewer-classes. Finally, we present additional ablation results in Appendix D, and additional qualitative results in Appendix F.

	cKNN	AP	SemiPriori	SemiCL	All	Known	New
(1)	✗	✗	✓	✓	$60.1^{-2.8}$	$70.1^{+5.7}$	$55.1^{-7.0}$
(2)	✓	✓	✓	✗	$61.7^{-1.2}$	$63.6^{-0.8}$	$60.7^{-1.4}$
(3)	✓	✓	✗	✓	$57.3^{-5.6}$	$61.8^{-2.6}$	$55.1^{-7.0}$
(4)	✓	✗	✓	✓	$54.6^{-8.3}$	$65.5^{+1.1}$	$49.1^{-13.0}$
(5)	✓	✓	✓	✓	62.9	64.4	62.1

Table 3. **Ablation study on effectiveness of SemiAG in CAL stage on CUB-200 [42] dataset.** Here, **cKNN:** consensus KNN graph; **AP:** affinity propagation; **SemiPriori:** semi-supervised prior knowledge; **SemiCL:** semi-supervised contrastive loss in projected feature space on `[CLS]` and `[P]`. Scores reported in clustering accuracy. Each proposed component favorably contributes to the overall performance.

(a) (b) (c) (d)

Figure 4. **The t-SNE [39] visualization of ViT embeddings on CIFAR-10 test set**. (a) is `[CLS]` embeddings from naive VPT model; (b) denotes our PromptCAL `[CLS]` embeddings; (c) denotes our PromptCAL ensembled `[P]` embeddings; (d) represents embeddings of an arbitraty PromptCAL unsupervised prompt. All figures share the same axis scale. The complete visualization is presented in Appendix D.

Effectiveness of contrastive affinity learning. As mentioned in Sec. 4.4, SemiAG dominates the large improvements of PromptCAL. First, we conduct ablation experiments on SemiAG in CAL stage, in Table 3. The 1^{st} row denotes the performance of using naive KNN with SemiPriori for pseudo labeling at CAL stage; while, the last row represents our full SemiAG setup. The 2^{nd}, 3^{rd}, and 4^{th} row represent PromptCAL without semi-supervised contrastive loss, semi-supervised prior knowledge (Sec. 3.3), and affinity propagation (Sec. 3.3), respectively. From the results, we can observe that incorporating each component has a clear contribution: `(a)` Naive KNN severely overfits `Known` and performs poorer (with nearly 2.8% and 7.0% accuracy drops on `All` and `New` classes, respectively) than SemiAG, due to its susceptibility to noisy neighborhoods. `(b)` Affinity propagation is the most consequential component (improving by 8.3% on `All` and 13% on `New`), which proves the importance of counteracting adverse effects of false negatives in contrastive loss by retrieving more reliable positives. `(c)` Retaining SemiCL is beneficial, which, we guess, is because it can push away noisy pseudo positive samples and, thus, prevent overfitting and degenerated solu-

	Prompt	$\mathbf{L}^{\mathrm{P}}_{\mathrm{semi}}$	$\mathbf{L}^{\mathrm{P}}_{\mathrm{CAL}}$	CAL stage	All	Known	New
(1)	✗	✗	✗	✗	$51.3^{-11.6}$	$56.6^{-7.8}$	$48.7^{-12.4}$
(2)	✓	✗	✗	✗	$51.1^{-11.8}$	$55.4^{-9.0}$	$48.9^{-12.2}$
(3)	✓	✓	✗	✓	$61.6^{-1.3}$	$68.9^{+4.5}$	$58.0^{-4.1}$
(4)	✓	✗	✓	✓	$61.2^{-1.7}$	$65.2^{+0.8}$	$59.2^{-2.9}$
(5)	✗	✗	✗	✓	$60.3^{-2.6}$	$64.8^{+0.4}$	$58.0^{-4.1}$
(6)	✓	✓	✓	✓	62.9	64.4	62.1

Table 4. **Ablation study on effectiveness of prompt-related components on CUB-200 dataset.** Here, **Prompt:** prompt-adapted backbone; $L^{\mathbf{P}}_{\mathbf{semi}}$**:** semi-supervised contrastive loss on [P] prompts; $L^{\mathbf{P}}_{\mathbf{CAL}}$**:** CAL loss on [P]; **CAL stage:** second-stage training. Scores reported in clustering accuracy. Each component favorably contributes to the overall performance gain.

	C50-L10			C25-L50			C10-L50		
Method	All	Known	New	All	Known	New	All	Known	New
GCD [41]	60.2	68.9	55.8	56.8	67.6	55.0	48.3	65.1	47.3
ORCA (ResNet) [4]	39.4	55.1	31.2	37.0	64.1	31.7	30.1	64.3	27.1
ORCA† (ViT) [4]	60.3	66.0	55.3	58.2	**79.9**	57.5	51.7	78.0	50.2
PromptCAL-1st (Ours)	62.7	74.7	56.6	60.2	70.7	58.5	48.7	68.4	47.6
PromptCAL-2nd (Ours)	**68.9**	**77.5**	**64.7**	**65.7**	76.9	**63.8**	**53.2**	**79.3**	**51.7**

Table 5. **Ablation study on few-annotation GNCD on CIFAR-100 [25] dataset.** Digits following 'C' and 'L' stand for percentages of known classes and labeling ratios. †denotes adapted methods. Scores reported in accuracy.

tions. (d) SemiPriori further benefits overall performance by about 5.6% on All and 7% on New, which manifests the importance of incorporating the prior knowledge to calibrate pseudo labels.

Role of discriminative prompt regularization. Table 4 presents the ablation results for prompt-related components of PromptCAL. The 1^{st} and 2^{nd} rows denote the GCD baseline and our warmed-up PromptCAL-1^{st}. We note that visual prompts make no significant difference to the performance. However, we argue that it is due to lack of semantic discriminative supervision. Specifically, by observing PromptCAL without semantic discrimination supervision (3^{rd} row) and PromptCAL without sample discrimination supervision (4^{th} row), we can infer that both semantic discriminativeness and sample-wise discriminativeness are critical. Generally, lack of semantic discriminativeness will cause severe overfitting on Known classes. Furthermore, semantic prompt tuning is beneficial for discovering novel classes, since PromptCAL surpasses its counterpart without any prompt-related component (5^{th} row) on New by 2.6%. To summarize, semantically-aware DPR plays a positive and auxiliary role in facilitating semantic discriminativeness especially in categorizing novel classes. In fact, we conclude from additional ablations in Appendix D that the gains of prompts are more significant on larger datasets.

To vividly illustrate this point, we present the t-SNE [39] visualization results in Fig. 4 (see complete results in Appendix D.3). Here, we directly draw conclusions that (a) naive VPT causes overclustering problem and lacks semantically discriminativeness; (b) our proposed DPR supervision increases semantic discriminativeness of supervised and unsupervised prompts, which further enhances semantic signals of DPR loss and enables DPR and CAL to synergistically improve the overall performance. We present more discussions on this in Appendix D.3.

Towards few-annotation GNCD. We further evaluate our PromptCAL against other SOTA methods on more challenging few-annotation setups on CIFAR-100 dataset, *i.e.*, fewer known classes and lower labeling ratios. We consider three setups in Table 5: (1) C50-L10: 50% classes are known in which 10% samples are labeled; (2) C25-L50: 25% classes are known in which 50% samples are labeled; (3) C10-L50: 10% classes are known in which 50% samples are labeled. Since the few-annotation can incur more open-set noises, we set $K = 5$ for PromptCAL to increase robustness to noisy pseudo-labels.

From results in Table 5, we conclude that PromptCAL is robust to both low-labeling and few-class scenarios, outcompeting all SoTAs with large margins. Practically, it is more demanding for models to infer novel semantic clustering when fewer classes are known under semantic shifts. This explains the lower performance of all models in setup (3) than in setup (1). Compared with GCD and ORCA, our PromptCAL can learn semantically robust representation and consistently achieve high performance in all setups. ORCA (ViT) [4] achieves stronger performance than GCD; while, our PromptCAL can still outperform ORCA with clear margins in all setups. For example, PromptCAL surpasses ORCA (ViT) by $\sim 8\%$ on All accuracy in C50-L10 and C25-L50. We again observe that PromptCAL-2^{nd} learning contributes most to the overall performance, which again proves our proposed method can effectively calibrate the learned representation with remarkable gains on New classes. This capability best suits GNCD problem.

5. Conclusion

In this paper, we propose a two-stage framework, PromptCAL, to tackle challenging GNCD problem. After the warm-up stage of semi-supervised contrastive learning, we iteratively and simultaneously conduct contrastive affinity learning and discriminative prompt regularization to calibrate semantic representations. Specifically, at each iteration, we leverage discovered pseudo affinities on generated affinity graphs to guide optimization of the class token and to reinforce the semantic discriminativeness of prompts and our prompt-adapted ViT backbone. Extensive experiments on multiple generic and fine-grained benchmarks showcase that PromptCAL achieves state-of-the-art performance. Additional evidences illustrates that our discriminative prompt regularization and contrastive affinity learning objectives achieve a synergistic effect. Moreover, PromptCAL exhibits remarkable gains on few-class and low-label settings for categorizing novel classes.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 3, 7

[2] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000. 3

[3] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *arXiv preprint arXiv:2208.05688*, 2022. 1

[4] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021. 1, 2, 3, 5, 6, 7, 8

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 6

[7] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. *arXiv preprint arXiv:2106.03719*, 2021. 3

[8] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 3

[9] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021. 2, 5, 6, 7

[10] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 1, 3

[11] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *arXiv preprint arXiv:2002.05714*, 2020. 2, 3, 5, 6, 7

[12] 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 8401–8409, 2019. 1, 2

[13] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*, 2021. 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4, 5

[15] Samer Hijazi, Rishi Kumar, Chris Rowen, et al. Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA*, 9, 2015. 1

[16] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2785–2795, 2022. 2, 3

[17] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. 3, 5

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 2, 3, 4

[19] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 610–619, 2021. 2

[20] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. 3

[21] Salar Hosseini Khorasgani, Yuxuan Chen, and Florian Shkurti. Slic: Self-supervised learning with iterative clustering for human action videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16091–16101, 2022. 2, 3, 5

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 3, 4

[23] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. 2

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 8

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[27] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015. 4

[28] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2006. 5

[29] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021. 1

[30] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 3

[31] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 3

[32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[33] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018. 5

[34] Vittal Premachandran and Ramakrishna Kakarala. Consensus of k-nns for robust neighborhood selection on graph-based manifolds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1594–1601, 2013. 3, 5

[35] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2019. 3

[36] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020. 3

[37] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*, 2021. 2, 3

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7, 8

[40] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020. 1, 2, 3

[41] 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 7492–7501, 2022. 1, 2, 3, 4, 6, 7, 8

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4, 6, 7

[43] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4148–4157, 2018. 4

[44] MB Wright. Speeding up the hungarian algorithm. *Computers & Operations Research*, 17(1):95–96, 1990. 6

[45] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016. 2

[46] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 1

[47] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1, 3

[48] Xingwei Yang, Lakshman Prasad, and Longin Jan Latecki. Affinity learning with diffusion on tensor product graph. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):28–38, 2012. 5

[49] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 1

[50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[51] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34:22982–22994, 2021. 2, 3

[52] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051, 2021. 2, 3

[53] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10875, 2021. 2, 3, 5

[54] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9462–9470, 2021. 2

[55] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 3