PROMPTCAL: CONTRASTIVE AFFINITY LEARNING VIA AUXILIARY PROMPTS FOR GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in semi-supervised learning (SSL) have achieved remarkable success in learning with partially labeled in-distribution data. However, many existing SSL models fail to learn on unlabeled data sampled from novel semantic classes and thus rely on the closed-set assumption. In this work, we adopt the *open-set* SSL setting and target a pragmatic but under-explored Generalized Category Discovery (GCD) setting. The GCD setting aims to categorize unlabeled training data coming from known or unknown novel classes by leveraging the information in the labeled data. We propose a two-stage Contrastive Affinity Learning method with auxiliary visual Prompts, dubbed PromptCAL, to address this challenging problem. Our approach discovers reliable affinities between labeled and unlabelled samples to learn better clusters for both known and novel classes. Specifically, we first embed learnable visual prompts into a pre-trained visual transformer (ViT) backbone and supervise these prompts with an auxiliary loss to reinforce semantic discriminativeness and learn generalizable affinity relationships. Secondly, we propose an affinity-based contrastive loss based on an iterative semi-supervised affinity propagation process which can further enhance the benefits of prompt supervision. Extensive experimental evaluation demonstrates that our method is effective in discovering novel classes even with limited annotations and surpasses the current state-of-the-art on six benchmark datasets (with more than 10% on CUB and StanfordCars, and a significant margin on ImageNet-100). Our code and models will be publicly released.

1 INTRODUCTION

The deep neural networks have demonstrated favorable performance in the Semi-Supervised Learning (SSL) setting (Van Engelen & Hoos, 2020; Zhai et al., 2019; Xie et al., 2020; Hijazi et al., 2015). Some recent works can even achieve comparable performance to their fully-supervised counterparts using few annotations for image recognition (Tarvainen & Valpola, 2017; Xie et al., 2020; Cai et al., 2022). However, these approaches heavily rely on the *closed-world* assumption that unlabeled data share the same underlying class label space as the labeled data (Yang et al., 2021; Geng et al., 2020). In many realistic scenarios, this assumption does not hold true because of the dynamic nature of real-world tasks where novel classes can appear in addition to known classes. Furthermore, exhaustive annotations will incur exorbitant costs, and sometimes it is intractable to cover all classes.

In contrast to SSL, generalized category discovery (GCD) (Vaze et al., 2022) is a nascent but more pragmatic and challenging setting due to the *semantic shift*, *i.e.*, unlabeled data can be sampled from other unseen distributions (Yang et al., 2021). To be more specific, GCD intends to categorize the unlabeled data given the information of some labeled data, but the unlabeled data contains novel classes that are not included in the labeled dataset. The key challenge in GCD is to discriminate among novel classes when only the ground truth of known classes is given in the training set. Therefore, properly guiding the model to discover latent semantic information in the unlabeled set with novel classes is a crucial and non-trivial problem.

A recent seminal work, GCD (Vaze et al., 2022), takes advantage of the large-scale pre-trained visual transformer (ViT), and learns robust clusters for known and novel classes with a semi-supervised

contrastive learning phase on downstream target datasets. However, we discover that the remarkable potential of pre-trained ViT is actually suppressed by this practice, due to the inherent limitations of its self-supervised contrastive learning on the unlabeled set, *i.e.*, considering as false negatives different unlabeled samples from the same underlying class. Empirically, abundant false negatives in contrastive training can deteriorate the compactness and purity of semantic clustering (Huynh et al., 2022; Khorasgani et al., 2022). We argue that abundant reliable pseudo-positive pairs can be mined by iterative affinity learning and favorably facilitate semantic clustering.

Meanwhile, although a frozen pre-trained backbone in GCD can alleviate overfitting on known classes (Cao et al., 2021; Vaze et al., 2022), this constrains the network to adapt itself to downstream datasets, and to learn beneficial semantic discriminativeness. Although visual prompt tuning (VPT) for ViTs (Jia et al., 2022) manifests effectiveness in fully-supervised learning, we observe no superiority of VPT compared with fine-tuning the last block, especially on small datasets.

To address the above limitations, we propose Prompt-based Contrastive Affinity Learning (PromptCAL). Our approach aims to discover both old and novel semantic clusters in the unlabeled data by simultaneously learning discriminative prompts and better feature representations via reliable affinity information in sample neighborhoods. Our proposed Contrastive Affinity Learning (CAL) and discriminative Multi-Prompt Clustering (MPC) achieve a synergistic effect. *Firstly*, CAL selects abundant reliable pseudo-positive pairs based on the proposed reliable affinity graph, further enhancing the semantic discriminability of the prompts. *Secondly*, MPC refines ViT semantic representation for better pseudo-labeling on refined affinity graphs. In this process, as model representation is iteratively refined, we can obtain higher quality pseudo-positives for further self-training as well as obtaining higher-quality clusters from the unlabelled data.

Our contributions can be summarized in three folds: (1) We propose a two-stage contrastive affinity learning framework, **PromptCAL**, to solve the generalized category discovery problem. (2) We propose contrastive affinity learning that selects reliable pseudo-positive pairs by learning from constructed affinity graphs based on a semi-supervised affinity propagation strategy. (3) We demonstrate that our PromptCAL can achieve significantly better performance on multiple benchmarks compared with previous state-of-the-art, thereby showing its effectiveness.

2 METHOD

The challenging aspect of GCD in comparison to SSL setting is the requirement of clustering novel semantics under both semantic shifts and missing annotations (Yang et al., 2021). However, existing methods (Vaze et al., 2022; Han et al., 2020; Zhao & Han, 2021; Zhong et al., 2021) cannot reliably harness the massive pre-trained knowledge in self-supervised learning. Moreover, recent SoTAs (Vaze et al., 2022; Cao et al., 2021) lack suitable strategies to adapt the pre-trained backbone to learn discriminative semantic information without overfitting on known classes.

To this end, we propose a prompt-based contrastive affinity learning method, consisting of two synergistic learning objectives: multi-prompt clustering (MPC) and contrastive affinity learning (CAL) (the basic formulation in Sec. 2.1). Specifically, in the first stage, we learn warm-up representation (in Sec. 2.2) for further tuning. In the second stage for CAL, we discover reliable pseudo-positives on affinity graphs in the embedding space based on proposed semi-supervised affinity generation (SemiAG) (in Sec. 2.3). Next, we propose our task-aligned contrastive affinity loss (in Sec. 2.4) based on SemiAG and a graph sampling technique.

2.1 FORMULATION AND PRELIMINARIES

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

Problem Definition. Our setting follows GCD (Vaze et al., 2022). The task of Generalized Category Discovery (GCD) assumes 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 (Vaze et al., 2022), we assume the knowledge of $|\mathcal{C}|$ and access to a validation set where only samples from $\mathcal{Y}_l$ are labeled.

Notation. We take a self-supervised ImageNet pre-trained ViT, DINO (Caron et al., 2021), as our backbone (Sharir et al., 2021). Then, we denote the deep visual prompt-adapted ViT backbone (Jia et al., 2022) as $f(\cdot|\theta, \theta_{\text{P}})$ parameterized by prompts θ_{P} and last block weights θ. For each mini-batch $\mathcal{B}$, there are two augmented views for each sample. Given a sample vector $\boldsymbol{x} \in \mathcal{B}$, we can extract its embedding $\boldsymbol{h} = f(\boldsymbol{x}|\theta, \theta_{\text{P}}) \in \mathcal{H}$ and project $\boldsymbol{h}$ into feature vector $\boldsymbol{z} = g(\boldsymbol{h}|\theta_{\text{H}}) \in \mathcal{Z}$ 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 define an extended contrastive loss based on standard supervised contrastive loss (Khosla et al., 2020). Given a query vector $\boldsymbol{t}_q$ and a set of key vectors $\boldsymbol{T}_k$, we define:

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

where τ is the temperature parameter of contrastive loss, and $\cdot$ denotes the cosine similarity operation. Then $\mathcal{P}(\boldsymbol{t}_q)$ and $\mathcal{N}(\boldsymbol{t}_q)$ represent the positive and negative sets of $\boldsymbol{t}_q$, respectively. We define anchor set as $\mathcal{A}(\boldsymbol{t}_q) = \mathcal{P}(\boldsymbol{t}_q) \bigcup \mathcal{N}(\boldsymbol{x}_q)$. Here, $\boldsymbol{t}_q$ and elements of $\boldsymbol{T}_k$ are L_2 normalized.

2.2 Warm-up with Multi-prompt clustering

Although computation overhead gets reduced by only tuning the last block in (Vaze et al., 2022), it restricts the backbone from learning better semantically discriminative representations and adapting well to diverse downstream datasets, *e.g.*, low-resolution images or fine-grained datasets. Meanwhile, we discover that naively adapting backbone with visual prompts (VPT) (Jia et al., 2022) overfits small dataset (refer to ablations on CUB-200 in section 3.5).

Motivated by (Lee et al., 2015; Wang et al., 2018), we propose multi-prompt clustering (MPC) to regularize and force prompts to learn semantically discriminative features with a task-related auxiliary loss. Suppose there are N_{P} tunable prompts before each ViT block. We assign the first N_{CLU} input prompts, at the last ViT block, as [CLU] prompts which are supervised by a task-related clustering loss in both training stages. All the remaining prompts are automatically tuned with no supervision, which provides the backbone with extra adaptability.

Specifically, we average the L_2-normalized embeddings from all [CLU] into an ensembled embedding $\overline{\boldsymbol{h}}_{\text{CLU}}$ and project it through an auxiliary head into $\overline{\boldsymbol{z}}_{\text{CLU}}$. Eventually, we conduct MPC on $\overline{\boldsymbol{h}}_{\text{CLU}}$ or $\overline{\boldsymbol{z}}_{\text{CLU}}$ by the same task loss on $\boldsymbol{h}$ or $\boldsymbol{z}$, but weighted by a hyper-parameter γ.

To enhance the initial representation for contrastive affinity learning, we apply warm-up training to adapt the pre-trained backbone to downstream target datasets and to learn [CLS] and [CLU] for next-stage training. The overall training objective in this stage is formulated as:

$$L_1(\boldsymbol{x}) = L_{\text{semi}}^{\text{CLS}}(\boldsymbol{z}) + \gamma L_{\text{semi}}^{\text{CLU}}(\overline{\boldsymbol{z}}_{\text{CLU}}) \tag{2}$$

where $L_{\text{semi}}^{\text{CLS}}$ and $L_{\text{semi}}^{\text{CLU}}$ represent the semi-supervised contrastive loss (SemiCL) on [CLS] and its MPC counterpart on [CLU], respectively, and γ is the MPC weight. Further, based on extended contrastive loss (Eq. 1), the SemiCL for token $\boldsymbol{z}$ is written as:

$$L_{\text{semi}}^{\text{CLS}}(\boldsymbol{z}) = (1-\alpha) L_{\text{con}}\Big(\boldsymbol{z}, \boldsymbol{Z}_{\mathcal{B}}; \tau, \mathcal{P}_{\text{self}}, \mathcal{N}_{\text{self}}\Big) + \alpha L_{\text{con}}\Big(\boldsymbol{z}, \boldsymbol{Z}_{\mathcal{B}_l}; \tau, \mathcal{P}_{\text{sup}}, \mathcal{N}_{\text{sup}}\Big) \mathbb{I}\Big(\boldsymbol{x} \in \mathcal{B}_l\Big) \tag{3}$$

where $\mathbb{I}$ is an indicator function. The first and second terms denote supervised and self-supervised contrastive loss on projected features of all and only labeled batches $\boldsymbol{Z}_{\mathcal{B}}$ and $\boldsymbol{Z}_{\mathcal{B}_l}$, respectively. $\mathcal{P}_{\text{self}}(\boldsymbol{z})$ and $\mathcal{P}_{\text{sup}}(\boldsymbol{z})$ respectively denote the embeddings of augmented counterparts and samples in same class, while $\mathcal{N}_{\text{self}}(\boldsymbol{z})$ and $\mathcal{N}_{\text{sup}}(\boldsymbol{z})$ denote the negative samples. Similarly, we can define the variables in the MPC loss function $L_{\text{semi}}^{\text{CLU}}$.

2.3 Semi-supervised Affinity Generation

Pseudo-labeling techniques, *e.g.*, nearest neighbors or pair-wise predictions as positives (Zhong et al., 2021; Cao et al., 2021; Fini et al., 2021; Han et al., 2020), in recent works are not robust to semantic shifts (Oliver et al., 2018). To address this issue, we propose a semi-supervised affinity generation (SemiAG) method with the assumption that local neighbors share the same semantics.

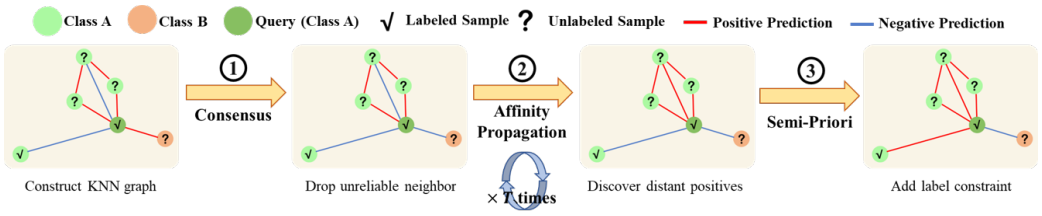

Figure 1: An intuitive example for SemiAG. Each step of the SemiAG process can either remove some false positives or retrieve new positives for the query embedding (in dark green). The distance between each pair is proportional to the cosine distance in the embedding space.

Specifically, we *first* construct an consensus affinity graph in $\mathcal{H}$ based on each neighborhood statistics (Premachandran & Kakarala, 2013). *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.

Consensus KNN graph. Given an embedding graph $G_{\mathcal{H}} = (\mathcal{V}, \mathcal{E})$ whose node set contains N_G embeddings as $\mathcal{V} = \{\boldsymbol{h_i}\}_{i=1}^{N_G}$ and edge set as $\mathcal{E} = \{e_{i,j} = \boldsymbol{h_i} \cdot \boldsymbol{h_j}\}_{i,j=1}^{N_G}$, we build a consensus graph $G_c = (g_{i,j})_{i,j=1}^{N_G}$ via consensus statistics. Each edge $g_{i,j}$ is defined as:

$$g_{i,j} = \begin{cases} |\{\boldsymbol{h_c} | \boldsymbol{h_i}, \boldsymbol{h_j} \in \mathcal{O}_K(h_c)), \forall \boldsymbol{h_c} \in \mathcal{V}\}| & i \neq j \\ 0 & i = j, \end{cases} \tag{4}$$

where, $\mathcal{O}_K(h_c) = \texttt{argtopK}_{\boldsymbol{h_j}}(\{\boldsymbol{h_j} \cdot \boldsymbol{h_c} | \boldsymbol{h_j} \in \mathcal{V}\})$ denotes the K-neighborhood of $\boldsymbol{h_c} \in \mathcal{V}$. Then, we can convert it into a probabilistic matrix $\tilde{G}_c$ by the row normalization.

Affinity propagation with priori. Furthermore, we leverage the graph diffusion (Yang et al., 2012) on the probabilistic matrix $\tilde{G}_c$ to propagate local affinities along multi-hop paths to characterize higher-order structural information. Specifically, we apply TPG diffusion algorithm (Yang et al., 2012), which iteratively computes the diffused graph $\tilde{G}_d$ as:

$$\tilde{G}_d^{(t+1)} = \tilde{G}_c \tilde{G}_d^{(t)} \tilde{G}^T + I, \tag{5}$$

where I is an identity matrix, and T is the total diffusion step. $\tilde{G}_d^{(t)}$ denotes the t-th step diffused graph and set $\tilde{G}_d^{(0)} = \tilde{G}$. We denote the final diffused graph as $\tilde{G}_d$.

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

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

On generated binarized affinity graph G_b, all positive pairs are taken as pseudo-labels for further training. Note that we compute the binarized graphs for both [CLS] and [CLU].

2.4 Contrastive Affinity Learning

In this subsection, we show how to use the generated pseudo labels to update the models.

Graph sampling with memory. One practical issue arises: 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 (Iscen et al., 2019). To strike a balance between the graph size and computation resources, we dynamically construct a sub-graph $G'_{\mathcal{H}}$ sub-sampled from

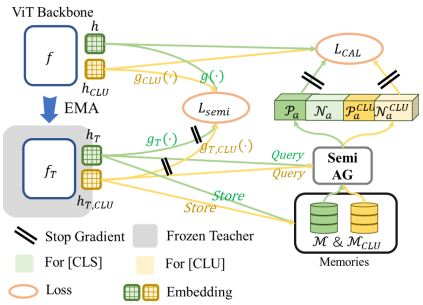

Figure 2: The overall framework of contrastive affinity learning. $g(\cdot)$, $g_{\text{CLU}}(\cdot)$, $g_{\text{T}}(\cdot)$, $g_{\text{T,CLU}}(\cdot)$ denote the [CLS] projection head, [CLU] projection head of the student and those of teacher.

the entire graph $G_{\mathcal{H}}$ with an extra embedding memory bank $\mathcal{M}$ and an exponentially moving averaged (EMA) teacher $f_{\mathrm{T}}, g_{\mathrm{T}}$, like MoCo (He et al., 2020). In each batch, the EMA teacher produces stable embeddings, which are enqueued to the first-in-first-out memory. The sub-graph $G'_{\mathcal{H}}$ is constructed by the embeddings in the memory and teacher embeddings in the current batch as $\boldsymbol{Z}_{\boldsymbol{T},\boldsymbol{\mathcal{B}}} = \mathcal{M} \bigcup \{\boldsymbol{h}_{\mathrm{T}} = f_{\mathrm{T}}(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{B}\}$. In this way, we can apply SemiAG on the sub-graph on the fly with adjustable memories.

Contrastive affinity loss. Given the sub-graph $G'_{\mathcal{H}}$ and its corresponding binarized graph G'_b by SemiAG, we formulate a contrastive affinity loss to refine the semantic representation as:

$$L_{\mathrm{CAL}}^{\mathrm{CLS}}(\boldsymbol{h}_i, G'_b) = L_{\mathrm{con}}(\boldsymbol{h}_i, \mathcal{M} \cup \boldsymbol{Z}_{\mathrm{T},\boldsymbol{\mathcal{B}}}, \tau_a, \mathcal{P}_a(\boldsymbol{h_i}), \mathcal{N}_a(\boldsymbol{h_i})) \tag{7}$$

where τ_a is a temperature parameter, and the positive set $\mathcal{P}_a$ contains the teacher embedding of the generated counterparts and the embedding of samples with the same pseudo label in the memory bank. Different from SemiCL in Eq. 3, CAL loss focuses on embedding space $\mathcal{H}$ rather than feature space $\mathcal{Z}$. But both of them are conducted on both [CLU] and [CLS] tokens.

Besides, we also apply the MPC loss in Eq. 3 in this contrastive affinity learning process to enhance the representation ability of prompts. To increase the consistency between the teacher and student, we modify the feature space from the all batch $\boldsymbol{Z}_{\boldsymbol{\mathcal{B}}}$ into the teacher and memory space $\boldsymbol{Z}_{\boldsymbol{T},\boldsymbol{\mathcal{B}}}$. Please refer to the Appendix for more details.

3 Experiments

3.1 Datasets

We evaluate PromptCAL on six benchmarks, *i.e.*, CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), ImageNet-100 (Krizhevsky et al., 2017), CUB-200 (Wah et al., 2011), StandfordCars (Krause et al., 2013), and Aircraft (Maji et al., 2013). A summary of all datasets is listed in Appendix A. First, we randomly split each dataset by known/novel classes using pre-defined $|\mathcal{C}_{kwn}|$ (#Known Classes in Table 7). Then, for each sampled known class, we further randomly sample a pre-defined ratio of samples into $\mathcal{D}_l$. All unlabeled samples constitute $\mathcal{D}_u$. We set labeling ratio $\frac{|\mathcal{D}_l|}{|\mathcal{D}|} = 50\%$ for all known classes on all datasets for main comparison in Sec. 3.4, following (Vaze et al., 2022). We adopt the same dataset split of $\mathcal{D}_l$ as in (Vaze et al., 2022). Meanwhile, we also present results in more challenging GCD setting with smaller $|\mathcal{C}_{kwn}|$ and $\frac{|\mathcal{D}_l|}{|\mathcal{D}|}$ in Sec. 6.

3.2 Evaluation Protocol

We follow the evaluation protocol in GCD (Vaze et al., 2022). For all models except for ORCA (Cao et al., 2021), we perform SemiKMeans (Vaze et al., 2022) on the predicted embeddings to get the cluster assignment for each sample. Then, all cluster prototypes are mapped through the optimal assignment solved by Hungarian algorithm (Wright, 1990) to their ground-truth classes. We report the standard accuracy scores on *Known*, *Novel*, and *All* classes. For PromptCAL, we report scores with standard clustering accuracy on the predicted embeddings from the student model. Additionally, we adapt and reproduce ORCA (Cao et al., 2021) in our setting and conduct comparisons in section 3.5. Since ORCA is a classification model, we evaluate it by the classification accuracy on known classes, and clustering accuracy on novel and overall classes. More details about adapted ORCA are given in Appendix B.

3.3 Implementation Details

For our PromptCAL, we set $\alpha = 0.35$ and $\tau = 1.0$ following GCD (Vaze et al., 2022) for both stages. We freeze the first 11 blocks and only tune the last block together with learnable prompts. Regrading prompt tuning, we pre-append $N_{\mathrm{P}} = 5$ prompts at each layer and assign first $N_{\mathrm{CLU}} = 2$ out of 5 as [CLU] prompts in all of our experiments. In the second training stage, we initialize the ViT backbone and two projection heads with the warm-up representation, and set $\beta = 0.6$, $\tau_a = 0.07$, memory size $|\mathcal{M}_{\mathrm{CLU}}| = |\mathcal{M}| = 4096$, neighborhood size $K = |\mathcal{M}|/(2|\mathcal{C}|)$, negative sample number to be 1024 for all datasets, unless specified. For SemiAG parameters, we set

threshold q to be 80% quantile value over the above-average non-zero affinity distribution for all fine-grained datasets, and 50% for all generic datasets. We empirically find that setting step $T = 1$ in SemiAG works well in most cases. Our teacher model is initialized by the student weights at the beginning, and we conduct momentum updates with a momentum factor of 0.999 at each iteration. During the inference, the [CLS] representation of the student model is for prediction. More detailed implementation details and baseline setups are discussed in Appendix B.1.

3.4 Main Results

In this section, we extensively conduct the performance comparison of PromptCAL against our baseline (GCD), and several adapted SoTA in other related settings, on six benchmarks.

Evaluation on generic datasets. We first evaluate PromptCAL on three generic datasets, *i.e.*, CIFAR-10/100, and ImageNet-100 in Table 1. The results in Table 1 show that our PromptCAL achieves the state-of-the-art performance and significantly surpasses our baseline method GCD on all three datasets ($\sim 6\%$ on CIFAR-10, $\sim 6\%$ on CIFAR-100, and $\sim 8\%$ on ImageNet-100 for *All* classes improvements) as well as Novel Category Discovery (NCD) SoTA methods *i.e.*, UNO+ (Fini et al., 2021) and RankStats+ (Han et al., 2020)). Besides, compared to GCD, PromptCAL manifests its advantage over all methods on *New* classes without sacrificing the performance on *Known* classes, *i.e.*, about 1% on CIFAR-10, 10% improvement on CIFAR-100 and ImageNet-100.

Comparing the 1^{st} stage and 2^{nd} stage PromptCAL, we observe large performance boost. In addition, we also notice that both stages of PromptCAL have significant contributions to the final performance on the generic dataset, *e.g.*, the first stage improves $\sim 3\%$ and $\sim 2.6\%$ over GCD on CIFAR-100 and ImageNet-100, respectively; while the second stage further improves by 2.9% and 10%. To summarize, above results in Table 1 validate the effectiveness of PromptCAL on generic datasets.

Method	CIFAR-10			CIFAR-100			ImageNet-100		
Classes	All	Known	New	All	Known	New	All	Known	New
KMeans	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+	68.6	98.3	53.8	69.5	**80.6**	47.2	70.3	**95.0**	57.9
GCD	91.5	**97.9**	88.2	73.0	76.2	66.5	74.1	89.8	66.3
PromptCAL (1st Stage)	97.1	97.7	96.7	76.0	80.8	66.6	76.7	90.5	69.8
PromptCAL	**97.5**	97.3	**97.6**	**78.9**	80.3	**76.1**	**82.0**	93.7	**76.1**

Table 1: Evaluation on generic datasets. Scores reported in accuracy.

Evaluation on fine-grained datasets. We also report results on fine-grained datasets to demonstrate the PromptCAL effectiveness. The KMeans (Arthur & Vassilvitskii, 2006) represents running KMeans++ Algorithm on DINO representation, and thus its performance reflects the representation adaptability and learning difficulty of pre-trained DINO on downstream datasets. Apparently, compared with generic datasets in Table 2, category discovery on fine-grained datasets is more challenging. Despite the challenging nature, our PromptCAL consistently exceeds RankStats, UNO, and GCD on *All* and *New* classes by more than 10% overall accuracy on CUB-200 and StanfordCars.

Different from the results on generic datasets (in Table 1), experimental results on fine-grained datasets display that the highest performance gain of PromptCAL originates from the second stage training. Noticeably, the warm-up stage performance of PromptCAL even drops compared with GCD on CUB-200 and Aircraft-100; while, CAL in 2nd stage achieves consistent improvement on all datasets. These results validate our initial hypothesis and our motivation of proposing SemiAG and CAL, *i.e.*, properly generating reliable pseudo-labels for further training is crucial.

3.5 Ablation and analysis

In this section, we conduct extensive ablation experiments on challenging CUB-200 dataset to draw interesting conclusions which reveal the contributions of each PromptCAL component. Together with ablation results, we present some in-depth analysis on the superiority of SemiAG and respon-

Method	CUB-200			StanfordCars-196			Aircraft-100		
	All	Known	New	All	Known	New	All	Known	New
KMeans	34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8
RankStats+	33.3	51.6	24.2	28.3	61.8	12.1	27.9	55.8	12.8
UNO+	35.1	49.0	28.1	35.5	**70.5**	18.6	28.3	**53.7**	14.7
GCD	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9
PromptCAL (1st Stage)	51.1	55.4	48.9	42.6	62.8	32.9	44.5	44.6	44.5
PromptCAL	**62.1**	**66.0**	**60.1**	**51.9**	69.4	**43.5**	**48.3**	47.9	**48.5**

Table 2: Evaluation on fine-grained datasets. Scores reported in accuracy.

sibility of visual prompts in PromptCAL. Further, we explore how PromptCAL performs in more challenging scenarios of decreased labeling ratios and increased openess, *i.e.*, more classes are unlabeled and novel, which are closer to real-world applications. Finally, we analyze the effect and sensitivity of hyper-parameters of PromptCAL in Appendix D, and present some visualization results in Appendix D.3.

cKNN	AP	SemiPriori	SemiCL	All	Known	New
✗	✗	✗	✗	56.2	65.4	51.6
✓	✓	✓	✗	60.5	65.0	58.3
✓	✓	✗	✓	61.3	67.5	58.2
✓	✗	✓	✓	57.0	64.1	53.4
✓	✓	✓	✓	62.1	66.0	60.1

Table 3: Ablation study on effectiveness of SemiAG in the second-stage on CUB-200 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 [CLU]. Scores reported in clustering accuracy. Each component favorably contributes to the overall performance.

Prompt	SemiCL (CLU)	$L_{\text{CAL}}^{\text{CLU}}$	CAL stage	All	Known	New
✗	✗	✗	✗	51.3	56.6	48.7
✓	✗	✗	✗	51.1	55.4	48.9
✓	✓	✗	✓	59.4	69.2	54.6
✓	✗	✓	✓	60.2	65.1	57.8
✗	✓	✓	✓	61.3	68.3	57.8
✓	✓	✓	✓	62.1	66.0	60.1

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

Effectiveness of contrastive affinity learning. As mentioned in main results (section 3.4), the SemiAG dominates the large improvements of PromptCAl on most datasets. We begin by presenting the ablation experiments related to contrastive affinity learning in the second stage in Table 3.

In the table, the 1st row denotes the performance of using naive KNN as pseudo-positives for CAL loss; while, the last row represents our full PromptCAL setup. The 2nd, 3rd, and 4th row signify PromptCAL without affinity propagation (section 2.3), semi-supervised prior knowledge (section 2.3), and semi-supervised contrastive loss, respectively.

From Table 3, we can observe that incorporating each component has a clear contribution: (a) Affinity propagation is the most consequential to final performance (improving by 5.1% on *All* and 6.7% on *New*), which proves the importance of counteracting adverse effects of false negatives in contrastive loss by retrieving abundant reliable positives. (b) Naive KNN perform significantly poorer than SemiAG, due to its susceptibility to noisy neighborhoods. (c) Retaining semicl is beneficial, and we postulate it is because the self-supervised loss can push away noisy pseudo-positive samples as well as prevent overfitting on $\mathcal{D}_l$ or representation collapse. (d) SemiPriori further benefits overall performance by $\sim 1\%$ on *All* and 1.8% on *New*, which manifests the importance of incorporating the prior knowledge from $\mathcal{D}_l$ for better affinity learning.

We further provide figures of memory precision and recall of different psuedo-labeling strategies during contrastive affinity learning (in Fig. 3). We can observe that SemiAG and SemiAG w/o SemiPriori has balanced precision and recall; while, KNN and SemiAG w/o AP suffer from either low precision or low recall. We argue that both precision and recall matters in CAL method. Moreover, SemiAG has higher precision and recall than SemiPriori due to priori constraint (eq. 6).

Role of multi-prompt clustering. Table 4 presents the ablation results for prompt-related components in PromptCAL. The 1^{st} and 2^{nd} rows denote the GCD baseline and our warm-up model with [CLU] prompts. 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) underperforms Prompt-

CAL without sample discrimination supervision (4^{th} row) by 0.8% on All and 3.2% on New, we can infer that semantic discriminativeness is critical to learning novel classes. Furthermore, supervised visual prompts are beneficial to discovering novel classes, since PromptCAL surpasses its counterpart without prompts and related MPC loss (5^{th} row) on *New* by 2.3%. To summarize, MPC plays a beneficial and auxiliary role in learning about novel classes.

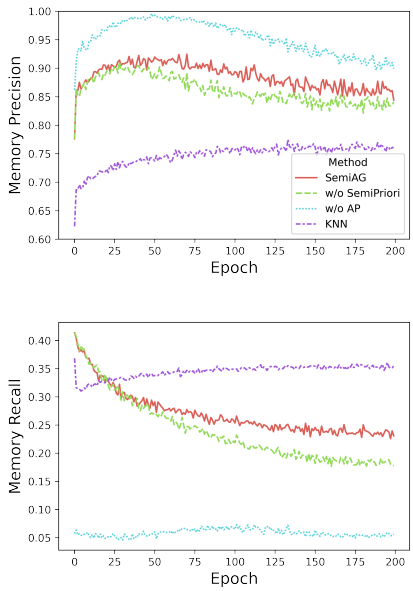

Figure 3: Memory precision and recall curves of SemiAG (red), SemiAG without SemiPriori (green), SemiAG without AP (blue), and naive KNN (purple), recorded at each training epoch. Corresponding to the 5^{th}, 3^{rd}, 4^{th}, and 1^{st} row in Table 3. For all setups, we keep all other hyperparameters fixed. Precision and recall are computed with ground-truth labels of embeddings in the [CLS] memory bank. The curves illustrate that both high precision and high recall matter in contrastive affinity learning.

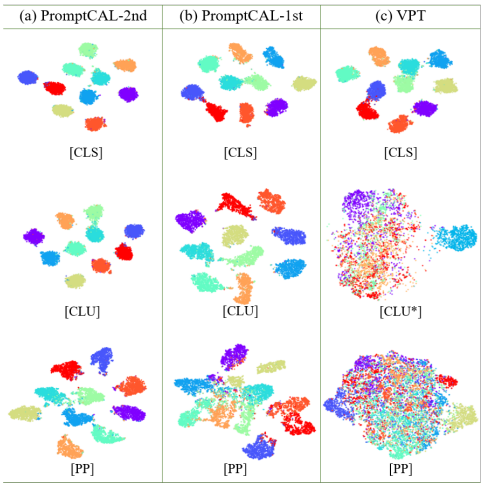

Figure 4: t-SNE (Van der Maaten & Hinton, 2008) visualization of ViT embeddings on CIFAR-10 test set for PromptCAL at the 2nd stage (left), PromptCAL at the 1st stage (middle), and warm-up training with naive VPT without MPC (right). Here, [CLS], [CLU], [CLU]*, and [PP] denote embeddings from ViT class token, our defined clustering prompts, ensembled unsupervised prompts, and a randomly picked and automatically learned prompt without supervision (for PromptCAL stage-2, it is unsupervised in both stages), respectively. The clustering quality of embeddings shows that MPC reinforces the semantic discriminativeness of [CLU], same holds for [PP] despite no supervision. Each component favorably contributes to the overall performance.

To vividly illustrate this point, we present the t-SNE (Van der Maaten & Hinton, 2008) visualization results in Fig. 4. Comparing VPT without MPC with PromptCAL at 1^{st} stage, we can observe that (a) CLU supervised by MPC can learn more semantically meaningful embeddings, in sharp contrast to automatically learned prompts in VPT. (b) Though not supervised, automatically learned prompts PP in PromptCAL 1^{st} learn semantically enhanced structures, benefiting from MPC on CLU. Moreover, by comparing PromtCAL at 2^{nd} stage and PromptCAL 1^{st}, we can conclude that: (c) [CLU] supervised by CAL loss can learn better semantic clustering than those supervised by SemiCL, and better benefit [PP] prompts. (d) [CLS] of PromptCAL 2^{nd} learns more compact and better-separated clusters compared with that of PromptCAL 1^{st}. Therefore, we can conclude that the second stage enhances the prompts' potential using CAL loss, which further enables prompts and CAL to synergically improve the overall performance.

Comparing with other related SoTA. In Table 5, we compare our PromptCAL with the recent SoTA on CIFAR-10/100 (Krizhevsky et al., 2009) dataset in another related setting, *i.e.*, open-world semi-supervised learning. Here, we report reproduced results of the original ORCA (with ResNet). We also adapt ORCA with a pre-trained DINO backbone and only tune the last block, denoted as ORCA (DINO), to make the comparison fair. *Firstly*, the performance of our reproduced ORCA (DINO) surpasses its original version by a large margin on both datasets, which manifests the benefits of using large-scale pre-trained ViT. *Secondly*, our PromptCAL outperforms the competitive ORCA (DINO) on both datasets with notable advantages on *New* classes on CIFAR-100 with 23%.

Method	CIFAR-10			CIFAR-100		
Classes	All	Known	New	All	Known	New
KMeans	83.6	85.7	82.5	52.0	52.2	50.8
GCD	91.5	**97.9**	88.2	73.0	76.2	66.5
ORCA* (ResNet)	89.4	87.9	90.2	54.9	66.3	43.3
ORCA* (DINO)	96.5	96.3	96.6	67.3	**83.9**	53.1
PromptCAL (1st Stage)	97.1	97.7	96.7	76.0	80.8	66.6
PromptCAL	**97.5**	97.3	**97.6**	**78.9**	80.3	**76.1**

Table 5: Evaluation on generic datasets. Scores report in accuracy. Asterisk (*) denotes that we adapt the method on our dataset split.

Method	$\lvert\mathcal{C}_{kwn}\rvert = 50, \frac{\lvert\mathcal{D}_l\rvert}{\lvert\mathcal{D}\rvert} = 0.1$			$\lvert\mathcal{C}_{kwn}\rvert = 25, \frac{\lvert\mathcal{D}_l\rvert}{\lvert\mathcal{D}\rvert} = 0.5$			$\lvert\mathcal{C}_{kwn}\rvert = 10, \frac{\lvert\mathcal{D}_l\rvert}{\lvert\mathcal{D}\rvert} = 0.5$		
	All	Known	New	All	Known	New	All	Known	New
GCD	60.2	68.9	55.8	56.8	67.6	55.0	48.3	65.1	47.3
ORCA* (ResNet)	39.4	55.1	31.2	37.0	64.1	31.7	30.1	64.3	27.1
ORCA† (DINO)	41.8	**80.1**	49.8	23.0	**87.2**	25.3	9.7	**92.6**	10.8
PromptCAL (1st Stage)	62.7	74.7	56.6	60.2	70.7	58.5	48.7	68.4	47.6
PromptCAL	**68.9**	77.5	**64.7**	**65.7**	76.9	**63.8**	**53.2**	79.3	**51.7**

Table 6: Ablation study on few-annotation GCD on CIFAR-100 dataset. Scores reported in accuracy.

Compared with GCD and ORCA, we conclude that the second-stage training of PromptCAL exhibits great superiority in discovering novel classes.

Toward few-annotation GCD. To further evaluate our PromptCAL against three state-of-the-art counterparts, GCD, ORCA (ResNet), and ViT-adapted ORCA, on more challenging few-annotation setups on CIFAR-100 dataset, *i.e.*, fewer classes and samples are annotated. We here consider three setups in Table 6: (1) known class number $|\mathcal{C}_{kwn}| = 50$ with labeling ratio equals to 10%; (2) $|\mathcal{C}_{kwn}| = 25$ with labeling ratio equals to 50%; (3) $|\mathcal{C}_{kwn}| = 10$ with labeling ratios equals to 50%.

From the Table 6, PromptCAL is robust to both low-labeling ratios and few-class scenarios, outcompeting all SoTA methods with huge margins. Practically, it posits more challenges to models to infer new semantic clustering when the known class number deceases due to the semantic shift issue. However, PromptCAL is still able to maitain high performance, achieving 51.7% accuracy on new classes and surpassing GCD with $\sim 5\%$ on overall accuracy. Besides, we realize that the classification-based model, ORCA, achieves much poorer in few-class scenarios, which, we speculate, is because classfication model cannot effectively mine new semantic information and severely overfit on *Known* classes. In contrast, clustering models, GCD and our PromptCAL, can learn generalizable representation through the contrastive loss during training. This inherent representation learning behavior empower the network in generalized category discovery. Since our proposed SemiAG in PromptCAL can effectively calibrate the learned representation with an affinity self-learning process, it gains larger edges oon *New* classes.

4 Conclusion

In this paper, we propose a two-stage Contrastive Affinity Learning method with auxiliary visual Prompts for generalized category discover. Specifically, we first embed learnable visual prompts to reinforce semantic discriminativeness of method. Then we apply an iterative semi-supervised affinity propagation process to mine positive samples of both prompts and class tokens and optimize them with an affinity-based contrastive loss. Experimentally, we evaluated our method on six generalized category discover benchmarks and demonstrated its superiority with significant improvement, *e.g.*, $+10\%$ on CUB and StanfordCarson CUB and StanfordCars and a large margin on ImageNet100.

Reproducibility Statement

We introduce the datasets, experimental settings, and the details of our method on Section 3.1, 3.2, 3.3 and Appendix B and will make the code available after publication.

REFERENCES

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

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.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021.

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*, pp. 9650–9660, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

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*, pp. 9284–9292, 2021.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

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.

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*, pp. 8401–8409, 2019.

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.

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*, pp. 9729–9738, 2020.

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.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*, 2017.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. *arXiv preprint arXiv:1901.00544*, 2019.

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*, pp. 2785–2795, 2022.

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*, pp. 5070–5079, 2019.

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.

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*, pp. 16091–16101, 2022.

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.

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*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

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

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *EMNLP*, pp. 3045–3059, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *ACL*, pp. 4582–4597, 2021.

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.

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

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.

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*, pp. 1594–1601, 2013.

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.

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.

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

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

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pp. 268–285. Springer, 2020.

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*, pp. 7492–7501, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

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*, pp. 4148–4157, 2018.

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

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*, pp. 10687–10698, 2020.

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

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.

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*, pp. 1476–1485, 2019.

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.

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*, pp. 10867–10875, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

A Dataset Details

We evaluate PromptCAL on six benchmarks, *i.e.*, CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), ImageNet-100 (Krizhevsky et al., 2017), CUB-200 (Wah et al., 2011), StandfordCars (Krause et al., 2013), and Aircraft (Maji et al., 2013). The profile of six benchmark datasets is displayed in Table 7.

Dataset	CIFAR-10	CIFAR-100	ImageNet-100	CUB-200	Aircraft	StanfordCars
#Images in $\mathcal{D}$	50k	50k	127.2k	6k	6.6k	8.1k
#Classes	10	100	100	200	100	196
#Known Classes	5	80	50	100	50	98

Table 7: Benchmark datasets for our performance evaluation.

B Additional Implementation Details

B.1 Hyper-parameters, Architecture and Losses

We use a 12-layer base Vision Transformer (Sharir et al., 2021) with a patch size of 16 (ViT-B/16) as our backbone in all of our experiments. The backbone weight is initialized with pre-trained DINO (Caron et al., 2021) on the ImageNet-1K dataset. We use two randomly initialized DINO heads for the projection on [CLS] and [CLU], separately. In both stages, we train ViT with a batch size of

128 for all datasets and with standard SGD with a momentum of 0.9, a weight decay of 5×10^{-5}, and an initial learning rate of 0.1.

In the 1^st stage training of PromptCAL, we adopt an unsupervised knowledge distillation loss with a loss weight of 0.5 (which gradually decays to zero in the first 5 training epochs) on ImageNet-1K. We empirically find this loss has no significant effect on the final performance. Considering the potential effect of randomly initialized visual prompts, we add this loss.

To adapt ORCA (Cao et al., 2021) to our setting, we run ORCA (ResNet) on our dataset splits without any modification. For ORCA (DINO), we replace their ResNet backbone with the same pre-trained DINO. We also freeze the first 11 layers of DINO, and use the same optimizer, training schedules for ORCA (DINO), following the practice of PromptCAL and GCD.

For GCD (Vaze et al., 2022), we rigorously follow its original setup. Same as (Vaze et al., 2022), our data augmentation strategy includes: resizing to 224×224, random horizontal and vertical flipping, and color jittering. Moreover, we adopt the checkpoint with the best *Acc* on *Known* classes on the validation set in the first training stage, following GCd (Vaze et al., 2022); for the second stage, we use best CVI score on the validation set for best checkpointing, following (Han et al., 2019).

C Related work

C.1 Category discovery

Here, we first distinguish Genearlized Category Discovery (GCD) (Vaze et al., 2022) from Novel Category Discovery (NCD) (Zhong et al., 2021; Hsu et al., 2017; **?**) and Semi-Supervised Learning (SSL) (Van Engelen & Hoos, 2020). NCD aims to categorize the unlabeled samples coming from novel classes given some labeled data from known classes; while SSL focuses on learning with in-distribution unlabeled data. Differently, GCD more challenging and general in that it further looses the constraint and assumes that unlabeled data can come from both known and novel classes (Vaze et al., 2022). Most methods for category discovery focus on learning the robust representation under semantic and distribution shifts. For example, Earlier work formulate category discovery into deep transfer learning (Han et al., 2019; Hsu et al., 2019). Recently, self-supervised learning based methods show great success, such as SCAN (Van Gansbeke et al., 2020) and ORCA (Cao et al., 2021). SCAN (Van Gansbeke et al., 2020) first proposes to leverage the self-supervised pre-trained representation for unsupervised category discovery. ORCA (Cao et al., 2021) utilizes contrastive learning (Chen et al., 2020) pretrained model for knowledge transfer. The most related work to ours is GCD (Vaze et al., 2022), which utilizes semi-supervised contrastive learning to learn robust semantic clustering using large-scale pre-trained ViT. However, the remarkable potential of the pre-trained model is actually suppressed by this practice, since it only applies self-supervised contrastive learning on the unlabeled set to update weights. Unlike previous work, our PromptCAL simultaneously learns discriminative prompts and better feature representations via reliable affinity information in sample neighborhoods.

C.2 Visual prompt learning

Prompt learning originates from the field of NLP (Liu et al., 2021), which focuses on the efficient adaptation of the large-scale pre-trained language model. It is a competitive product of the "pre-training, fine-tuning" framework, by introducing a textual prompt to reformulate the downstream tasks into the form of the pretrained task (Lester et al., 2021; Li & Liang, 2021). It shows strong potential in data-efficient learning (Zhou et al., 2021; Gao et al., 2021) and attracts wide attention from the field of visual understanding recently (Jia et al., 2022). Visual prompt learning (VPT) (Jia et al., 2022) is the first to validate the effectiveness and potential of prompt tuning in the field of computer vision. Specifically, it adds some extra tokens in each block as prompts and tunes these prompts with downstream objectives while freezing the original pretrained models. VPT shows improved performance on general image classification tasks than entire finetuning and only adapting the last block, but its advantage seems not remarkable on small datasets. However, based on our experimental discoveries, we found no significant benefits brought by GCD in category discovery problem. Alternatively, we intend to propose the multi-prompt clustering learning to refine the visual prompt representation which indirectly benefits discriminative semantic clustering. Thus, our work differs GCD in different contexts and the learning goal.

C.3 POSITIVE MINING

Previous researches prove the effectiveness of pseudo-labeling in the open-set setting (Cao et al., 2021; Han et al., 2020; Van Gansbeke et al., 2020; Fini et al., 2021; Zhao & Han, 2021; Zhong et al., 2021). Many existing methods in related settings propose to learn pairwise similarity from labeled data to generate pseudo-labels on unlabeled data (Cao et al., 2021; Han et al., 2020; Van Gansbeke et al., 2020). MCL (Hsu et al., 2019) first proposes to utilize pairwise labels instead of class labels to solve the multi-class classification problem in various settings. Meanwhile, there is a line of recent works which utilize pseudo-labeling technique to discover positive pairs likely to share the same class. RankStats and RankStats++ (Han et al., 2020) propose to use ranking statistics to fine robust positive samples. NCL (Zhong et al., 2021) utilize k-Nearest Neighbors for each sample. UNO (Fini et al., 2021) formulates NCD as a classification problem and generates pseudo-labels by cross-view predictions. Our method differs from previous work in proposing a reliable pseudo-labeling method based on affinity graph for contrastive learning. We notice that (Iscen et al., 2019) proposes to utilize a graph diffusion method to solve the semi-supervised classification problem. But, there are several major differences with our work: first, our setting is open-set and thus more challenging than semi-supervised learning; second, we conduct efficient diffusion per iteration supported by a memory bank, while they conduct diffusion per epoch on all data; third, we compute affinity propagation on the affinity graph based on consensus information and semi-supervised priori, while they conduct propagation on naive KNN graph. Consensus KNN is first proposed in (Premachandran & Kakarala, 2013). However, our work is more related to deep clustering (Han et al., 2019), while their method is based on features extracted by conventional algorithms; we empower the consensus KNN with contrastive affinity learning which iteratively refines the representation graph and increases the neighborhood reliability; lastly, we propose multi-prompt clustering to learn semantic discriminative features and make affinity learning mutually benefit from the prompt clustering.

D MORE EXPERIMENT RESULTS

In this section, we present more experimental results on PromptCAL ablation studies.

D.1 ANALYSIS ON HYPER-PARAMETERS.

In PromptCAL, there exist three tunable hyperparameters: neighborhood size K for graph construction, quantile threshold value q in SemiAG, and the weight β of contrastive affinity loss. To investigate the sensitiveness of these parameters, we conduct ablation experiments on CIFAR-100 dataset by sampling: (1) the neighborhood $K = 5, 10, 20, 30$, (2) the quantile threshold $q = 0.4, 0.5, 0.6$, and (3) the weight $\beta = 0.4, 0.6, 0.8$. The neighborhood size controls the locality of consensus information. Its ablation result on CIFAR-100 dataset is shown in Table 8, which demonstrates that PromptCAL is rather robust when K is small and larger neighborhoods can hurt performance due to more noisy pseudo-positives included in SemiAG. Table 9 exposes that PromptCAL is rather robust to variation in threshold q on *All* classes. As we can observe, the performance on *All* when $q = 0.4$ only slightly drops. Interestingly, the *New* performance consistently improves when q increases with a little sacrifice in *Known* performance, while $q = 0.5$ best balances between *Known* and *New* classes. We explain that higher q will increase the reliability of *New* pseudo-positives in SemiAG, at the cost of more reliable *Known* positives being taken as false negatives.

For the loss weights, its value reflects the relative intensity of supervised contrastive learning with generated affinities *w.r.t.*self-supervised contrastive learning for sample-wise discriminativeness. From Table 10, we can observe that PromptCAL is rather robust to the increase and decrease of the CAL loss weight β by 0.2; the variations of *All* performance have only minor changes (less than 1%). Meanwhile, a larger β improves the *Known* class performance, while a smaller β focus more on learning *New* class. Thus, we set $\beta = 0.6$, which entitles the model with balanced capability on both.

K	CIFAR-100 All	CIFAR-100 Known	CIFAR-100 New	Aircraft All	Aircraft Known	Aircraft New
5	80.9	**85.5**	71.7	51.2	59.2	47.2
10	**81.2**	84.2	75.3	51.3	52.6	50.6
20	78.9	80.3	**76.1**	48.3	47.9	48.5

Table 8: Ablation study on the neighborhood size on the CIFAR-100 dataset.

Method	All	Known	New
GCD	51.3	56.6	48.7
PromptCAL-1st (MPC-2-3)	51.1	55.4	**48.9**
PromptCAL-1st (MPC-1-4)	**51.7**	**57.2**	**48.9**
PromptCAL-1st (MPC-5-0)	50.9	55.6	48.6
PromptCAL-1st (MPC-2-3, all-frozen)	51.1	55.4	**48.9**
PromptCAL-1st (MPC-2-3, no INKD)	51.1	56.3	48.5

Table 12: Superiority of visual prompts on CUB-200 dataset. Tuning more ViT blocks can lead to overfitting on *Known* classes.

q	All	Known	New
0.4	79.7	**83.4**	72.4
0.5	78.9	80.3	76.1
0.6	**80.3**	81.4	**77.9**

Table 9: Ablation study on the neighborhood size K on CIFAR-100 dataset.

β	All	Known	New
0.4	49.6	51.0	48.9
0.6	48.3	47.9	48.5
0.8	49.1	53.6	46.9

Table 10: Ablation study on the CAL loss weight β on Aircraft dataset.

Method	All	Known	New
GCD	73.0	76.2	66.5
GCD (tune 6 blocks)	62.9	74.5	39.8
VPT (5 prompts)	76.5	80.9	67.8
PromptCAL ($K = 10$)	**81.2**	**84.2**	**75.3**

Table 11: Superiority of visual prompts on CIFAR-100 dataset. Tuning more ViT blocks can lead to overfitting on *Known* classes.

D.2 Ablation study on tuning DINO.

To investigate the alternative architecture and testify to the effectiveness of prompt-adapted ViT backbone, we conduct an ablation study on DINO architecture on CIFAR-100 dataset in Table 12. By comparing 1^{st} and 2^{nd} row, we observe that tuning more ViT blocks actually hurts the overall performance and achieves extremely poor accuracy on *New* classes, which drops $\sim 27\%$ compared with GCD. This justifies that PromptCAL and GCD freeze the first 11 ViT blocks for training to avoid the loss of rich pre-trained knowledge. Moreover, we also conclude that the visual prompt-adapted ViT backbone significantly enhances the performance on *All*, *Known*, and *New*. One main possible reason is that CIFAR-100 is a low-resolution dataset and the backbone needs more powerful adaptability to downstream datasets and tasks.

D.3 Visualizations.

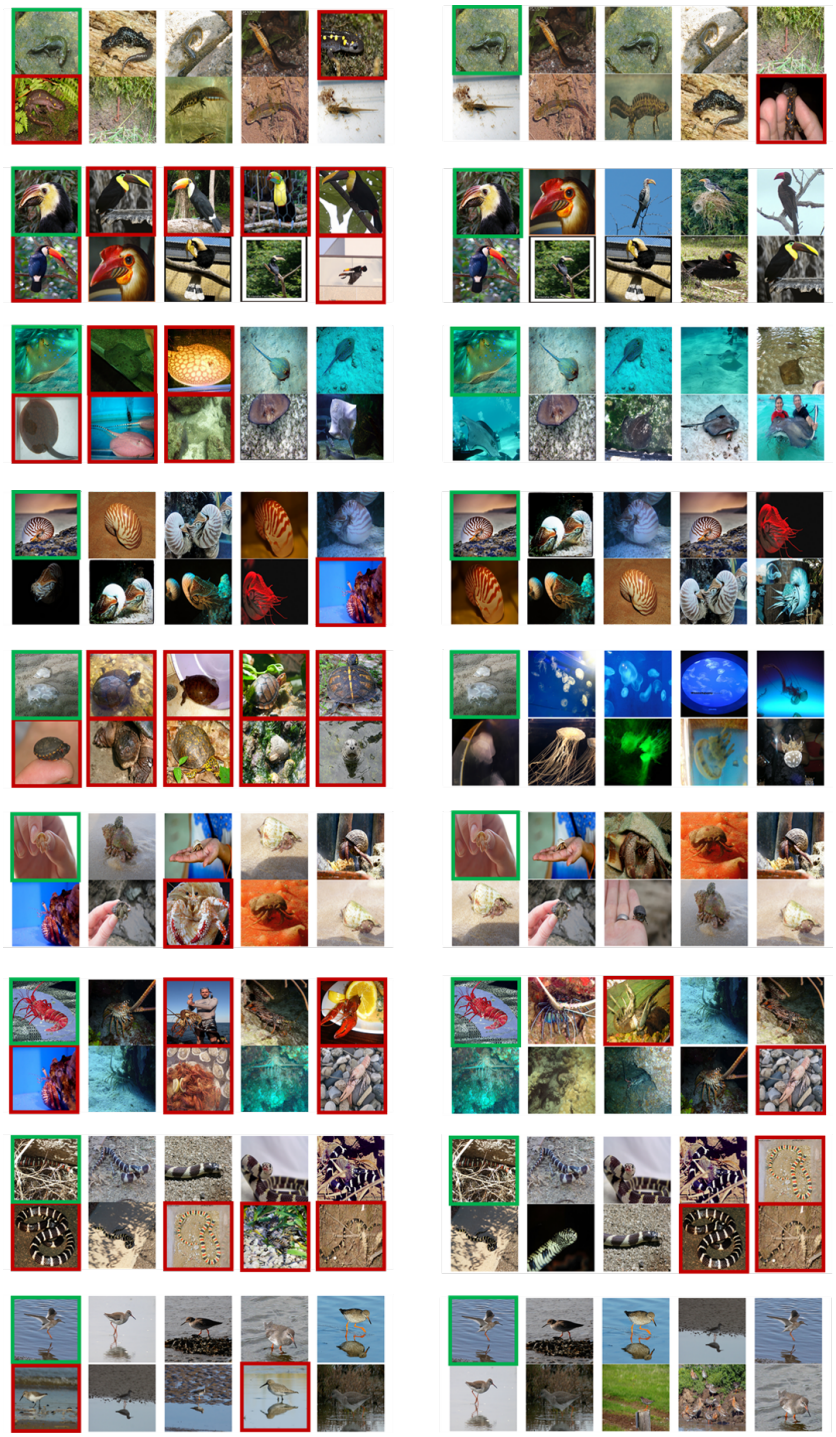

Figure 5: Visualization of the results on retrieving the 10 nearest neighbors of a given query image (with green border). Predictions of the first column is from GCD, and the second is from PromptCAL. The wrong predictions are marked with red borders; while, accurate predictions are not. We fisrt subsample 1000 images from ImageNet-100 conduct present results of 9 sampled examples, which contain both *New* and *Known* classes.