Information Fusion 119 (2025) 103020

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Mutual-support generalized category discovery

Yu Duan [a], Zhanxuan Hu [b], Rong Wang [d], Zhensheng Sun [e], Feiping Nie [c,*], Xuelong Li [d]

[a] School of Telecommunications Engineering, Xidian University, Xi'an 710071, PR China
[b] School of Information Science and Technology, Yunnan Normal University, Kunming, Yunnan, PR China
[c] School of Computer Science, School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), and the Key Laboratory of Intelligent Interaction and Applications (Ministry of Industry and Information Technology), Northwestern Polytechnical University, Xi'an, Shaanxi, PR China
[d] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), and the Key Laboratory of Intelligent Interaction and Applications (Ministry of Industry and Information Technology), Northwestern Polytechnical University, Xi'an, Shaanxi, PR China
[e] Zhijian Laboratory, Rocket Force University of Engineering, Xi'an, Shaanxi, PR China

ARTICLE INFO

Keywords:
Open-world semi-supervised learning
Mutual-support mechanism
Representation learning
Generalized category discovery

ABSTRACT

This work focuses on the problem of Generalized Category Discovery (GCD), a more realistic and challenging semi-supervised learning setting where unlabeled data may belong to either previously known or unseen categories. Recent advancements have demonstrated the efficacy of both pseudo-label-based parametric classification methods and representation-based non-parametric classification methods in tackling this problem. However, there exists a gap in the literature concerning the integration of their respective advantages. The former tends to be biased towards the 'Old' categories, making it easier to classify samples into the 'Old' groups. The latter cannot learn discriminative representations, decreasing the clustering performance. To this end, we propose Mutual-Support Generalized Category Discovery (MSGCD), a framework that unifies these two paradigms, leveraging their strengths in a mutually reinforcing manner. It simultaneously learns high-quality pseudo-labels and discriminative representations. It incorporates a novel *Mutual-Support mechanism* to facilitate symbiotic enhancement. Specifically, high-quality pseudo-labels furnish valuable weakly supervised information for learning discriminative representations, while discriminative representations enable the estimation of semantic similarity between samples, guiding the model in generating more reliable pseudo-labels. MSGCD is remarkably effective, achieving state-of-the-art results on several datasets. Moreover, *Mutual-Support mechanism* is not only effective in image classification tasks, but also provides intuition for cross-modal representation learning, open-world image segmentation, and recognition. The codes is available at https://github.com/DuannYu/MSGCD.

1. Introduction

Deep learning has achieved great success in numerous computer vision tasks [1–3]. This success is partially attributable to the existence of large labeled datasets and corresponding supervised training fashion. However, the acquisition of large labeled datasets proves challenging in many tasks and domains, demanding substantial annotation effort or domain expertise. Semi-supervised learning (SSL) [4–6] presents an alternative approach to alleviate the reliance on labeled data, training models using a limited number of labeled samples alongside a large number of unlabeled samples. Recent studies indicate that SSL can yield results comparable to supervised learning with only a modest number of annotations. Notably, these SSL approaches typically operate under the *closed-world* assumption, where both unlabeled and labeled training data share the same class label space [7]. This assumption constrains the applicability of SSL methods to scenarios where the class labels remain consistent across labeled and unlabeled data.

Recently, Generalized Category Discovery (GCD) [8], i.e., open-world semi-supervised learning [9], has been proposed to address a more realistic and challenging semi-supervised learning setting where unlabeled data can be from either previously known or new unseen categories. Approaches to GCD can be broadly categorized into two groups: pseudo-label-based parametric classification methods [9–11] and representation-based non-parametric classification methods [12–15]. The former constitutes a class of one-stage methodologies that leverage known annotation information and extrapolated pseudo-labels to train a parametric classifier capable of directly classifying unlabeled data. However, the latter encompasses a class of two-stage methods. These methods first learn a feature extraction network by integrating

* Corresponding author.
E-mail addresses: duanyuee@gmail.com (Y. Duan), zhanxuanhu@gmail.com (Z. Hu), wangrong07@tsinghua.org.cn (R. Wang), szs07@mails.tsinghua.edu.cn (Z. Sun), feipingnie@gmail.com (F. Nie), li@nwpu.edu.cn (X. Li).

https://doi.org/10.1016/j.inffus.2025.103020
Received 18 June 2024; Received in revised form 3 February 2025; Accepted 6 February 2025
Available online 14 February 2025

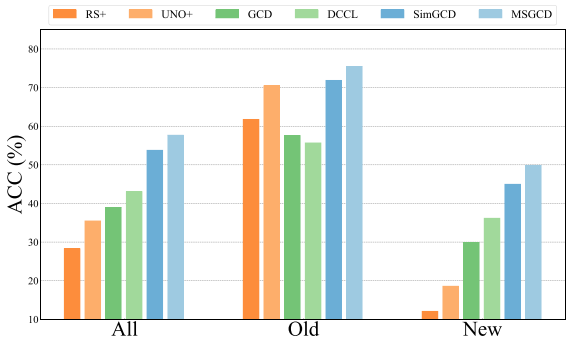

Fig. 1. Comparison between different kinds of GCD approaches on *Stanford Cars*, including pseudo-label-based parametric classification methods (orange), representation-based non-parametric classification methods (green), and joint training methods (blue). We divide the comparison results into three groups according to the categories of All, Old and New.

supervised and self-supervised representation learning. Subsequently, a non-parametric classifier, such as semi-supervised K-means++, is employed to derive the final classification results.

Recent studies validate that predictions made by parametric classification methods often manifest a bias towards 'Old' categories, while non-parametric classification methods adeptly categorize 'New' categories, as depicted in Fig. 1 and Table 2. This naturally raises a question: *Can the performance of GCD be further enhanced by integrating these two paradigms?* A recent state-of-the-art method ,SimGCD [16], demonstrated performance improvements through the integration of parametric classification and representation learning, as shown in Fig. 1. However, it directly incorporates classifier and projector into a Siamese network with two branches, thereby excluding the potential advantages of mutually exchanging valuable information. We argue that the dependable pseudo-labels generated by parametric classification methods can guide representation learning, and conversely, the representations conducive to clustering learned by non-parametric classification methods can offer meaningful semantic similarity information for parametric classification.

To address this objective, we propose Mutual-Support Generalized Category Discovery (MSGCD), an extension of SimGCD that introduces an innovative mutual-support mechanism, fostering collaboration between the classifier and projector for mutual benefit. In particular, MSGCD follows the architecture of SimGCD, employing a Siamese network with two branches to concurrently learn pseudo-labels for unlabeled samples and representations for all training samples. However, it distinguishes itself by incorporating an additional mechanism of information interaction designed to facilitate mutual support between these two branches. During training, MSGCD utilizes the pseudo-labels predicted by the classification branch as pseudo-supervision information to guide the representation learning branch. Moreover, it leverages discriminative representations to estimate semantic similarity information, thereby regularizing the outputs of the classification branch. Empirically, MSGCD consistently demonstrates improvements over SimGCD across various datasets. To the best our knowledge, MSGCD is the first GCD method that jointly learns representation and label prediction. The main contributions of this work are summarized as follows:

1. *New perspective.* To the best of our knowledge, this work is the first study to explore and validate the efficacy of integrating the advantages of parametric and non-parametric GCD methods.[1]
2. *New method.* We provide a novel *mutual-support mechanism* designed to facilitate information interaction between parametric and non-parametric GCD methods. Notably, this mechanism is independent of existing approaches and can be seamlessly integrated into them without the need for intricate modifications.
3. *New state-of-the-art results.* We conduct extensive visualization and ablation experiments to validate the efficacy of MSGCD. Furthermore, results from experiments on four standard datasets conclusively indicate that MSGCD outperforms existing state-of-the-art methods by a significant margin.

[1] Notably, the main motivation of SimGCD [16] is to build a simple baseline to parametric classifier rather than integrating the advantages of parametric classifiers and non-parametric classifiers.

2. Related works

2.1. Semi-supervised learning

The objective of semi-supervised learning (SSL) is to partition samples into distinct groups by utilizing a limited number of labeled data alongside numerous unlabeled samples [17,18]. The mainstream approaches in semi-supervised clustering are consistency-based methods, which have been extensively utilized across diverse domains. Broadly speaking, these methods primarily strive to guarantee consistent outputs from the model with various augmentations inputs. For instance, FixMatch [6], one of the extensively employed techniques, introduces consistency regularization by incorporating three augmentations. MixMatch [19] adopts a sharpened averaged prediction of multiple strongly augmented views as the pseudo label and enhances them by leveraging MixUp [20] trick. Furthermore, there exist several methods that enhance effectiveness from alternative perspectives. FreeMatch [21] dynamically adjusts class-specific confidence thresholds according to varying learning difficulties. CoMatch [22] and SimMatch [23] illustrate the advantages of self-supervised representation for semi-supervised learning tasks.

Despite the significant contributions of the aforementioned methods, they all operate under the assumption that each predefined category possesses labeled data. Nevertheless, satisfying this assumption proves challenging in numerous real-world tasks. Therefore, this paper addresses the more practical setting of open-set semi-supervised learning (SSL).

2.2. Novel category discovery

Novel Category Discovery (NCD) [24] focuses on discovering new categories in the unlabeled set by leveraging the knowledge learned from the labeled set, which brings close-set assumption SSL into a more realistic scenario. In typical NCD methods, a model is first trained on the labeled data and then utilized as an initialization for unsupervised clustering on the unlabeled data. Previous works [25,26] in this field utilize labeled data to train a binary classification model by leveraging pairwise image similarity. Subsequently, this trained binary classification model serves as a form of supervision for clustering on the unlabeled data. Recently, the increasing popularity of contrastive learning [27,28] has led to the proposal of numerous novel NCD methods. For instance, RankStat [29] proposes that self-supervised pretraining is advantageous for NCD. NCL [30] utilizes contrastive learning to improve representation learning. UNO [31] introduces a unified objective for concurrent learning on both unlabeled and labeled data. However, NCD assumes that the unlabeled data exclusively belong to known categories and cannot encompass data from novel categories during the training stage. It often proves challenging to meet these strict assumptions in real-world applications.

2.3. Generalized category discovery

Generalized Category Discovery (GCD) [8], also referred to as open-world semi-supervised learning [9], serves as an extension of traditional Semi-Supervised Learning (SSL) and Novel Category Discovery (NCD). In contrast to NCD, GCD permits unlabeled data to be associated with both known and novel classes. Recent developments in GCD primarily align with two trajectories: parametric classification methods based on pseudo-labels and non-parametric classification methods based on representations.

2.3.1. Parametric classification based on pseudo-labels

This category of methods strives to acquire a classifier capable of generating and utilizing pseudo-labels throughout the training process. TRSSL [11] proposes a pseudo-label method that is sensitive to class distribution, incorporating prior knowledge about class distribution during training. ORCA [9] introduces an extra classifier designed for novel categories, accompanied by an uncertainty adaptive margin mechanism to mitigate learning bias towards known categories. PIM [10] approaches GCD from an information-theoretic standpoint, presenting an objective function centered on constrained mutual information maximization.

However, these kind of methods prefer to group samples into 'Old' categories, because supervised information can directly guide the classifier's training. It will inevitably reduce the ability to discover 'New' categories.

2.3.2. Representation-based non-parametric classification methods

At the heart of this category of methods lies representation learning. Typically, these methods begin by extracting discriminative features through a sophisticated representation learning model. Subsequently, a fixed non-parametric classifier, such as semi-supervised K-means++, is employed to derive the final classification results. In the context of GCD [8], the proposal involves fine-tuning a self-supervised pre-trained model (DINO) through a combination of supervised and unsupervised contrastive learning. DCCL [13] improves GCD by introducing dynamic contrastive learning of conceptions, where visual conception estimation and learning of conceptional representation occur alternately. PromptCAL [32] adopts a two-stage contrastive affinity learning approach, introducing auxiliary visual prompts and alternately refining semantic prompts while conducting contrastive affinity learning. In addition, differing from the methods working on an assumption that the class number in the unlabeled data is known, PIM [10] and GPC [14] propose to estimate the class number automatically during training.

However, despite the promising outcomes achieved by non-parametric classification methods, they still suffer from the quadratic complexity of the clustering procedure, limiting their application on large-scale datasets. In addition, to our knowledge, MSGCD is the first to propose a cross-space joint training strategy. The basic idea of MSGCD can also be used in unsupervised and semi-supervised clustering tasks.

2.3.3. Joint-learning-based method

SimGCD [16] has recently integrated parametric classification and representation learning into a unified framework, showcasing simplicity alongside effectiveness and producing promising results across various benchmarks. However, SimGCD ignores the potential relationship between representation learning and outcome prediction. To handle these issues, we extend upon SimGCD, enhancing its capabilities by introducing a novel mutual support mechanism. This mechanism facilitates collaboration between parametric and non-parametric classification methods, thereby allowing them to mutually benefit from each other's strengths. In our previous work [33], we proposed an alternative method, called PCR, to jointly represent and learn the label space using EM optimization. To the best of our knowledge, MSGCD is the first method that jointly learns different spaces in an end-to-end manner.

3. Method

3.1. Problem formulation

GCD is a more challenging semi-supervised learning setting where an unlabeled dataset $\mathcal{D}^u = \{(\mathbf{x}_i^u, \mathbf{y}_i^u)\} \in \mathcal{X} \times \mathcal{Y}_u$ and a labeled dataset $\mathcal{D}^l = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\} \in \mathcal{X} \times \mathcal{Y}_l$ are provided, and $\mathcal{Y}_l \in \mathcal{Y}_u$. GCD aims to categorize novel categories in $\mathcal{D}^u$ by leveraging known prior from the labeled dataset $\mathcal{D}^l$. In addition, we follow existing studies [16] and assume that the number of labeled categories C_l in $\mathcal{Y}_l$ and unlabeled categories C_n in $\mathcal{Y}_u$ are known. K denotes the total number of categories.

3.2. Overview

To amalgamate the benefits of pseudo-label-based parametric classification methods and representation-based non-parametric classification methods, we introduce a novel framework named Mutual-Support Generalized Category Discovery (MSGCD). Illustrated in Fig. 2, MSGCD concurrently learns a label space and a representation space utilizing a Siamese network with two branches. Crucially, it incorporates a mutual-support mechanism facilitating mutual enhancement between these spaces. Formally, during training, given a mini-batch of data $B = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_b\}$, we partition it into two subsets: labeled set B^l and unlabeled set B^u. For each instance $\mathbf{x}_i$, we first generate two different views $(\mathbf{x}_i^a, \mathbf{x}_i^b)$ using random augmentation. Then, the features $(\mathbf{f}_i^a, \mathbf{f}_i^b)$ are generated using a shared backbone. Subsequently, Subsequently, these features are projected into a label space and a representation space, producing label predictions $(\mathbf{p}_i^a, \mathbf{p}_i^b)$ and ℓ_2-normalized representation vectors $(\mathbf{z}_i^a, \mathbf{z}_i^b)$, respectively. Finally, the label space and the representation space are jointly optimized. The overall objective of MSGCD is

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rep} + \mathcal{L}_{ms}\,, \tag{1}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{rep}$ represent the fundamental losses for label space learning and representation space learning, respectively. Additionally, $\mathcal{L}_{ms}$ constitutes an objective for our proposed mutual-support mechanism, comprising a graph-based distribution calibration loss and a prototype contrastive loss. In practical implementation, prevalent parametric/non-parametric classification methods, such as PIM [10], PromptCAL [32] and DCCL [13] can be employed to denote $\mathcal{L}_{cls}$ and $\mathcal{L}_{rep}$. For clarity, in this work, we build upon SimGCD [16], a straightforward yet effective baseline. Subsequently, we delve into the detailed explanation of these three components.

3.3. Label space learning

We follow adopt a prototypical classifier h for label space learning. Specifically, we randomly initialize K parametric prototype vectors $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K\}$, where $\mathbf{c}_k$ stands for the kth category. Correspondingly, the label prediction of x_i is $\mathbf{p}_i = \{p_i^1, p_i^2, \dots, p_i^K\}$, where

$$p_i^k = \frac{\exp\left(\frac{1}{\tau}(\mathbf{f}_i/\|\mathbf{f}_i\|)^T(\mathbf{c}_k/\|\mathbf{c}_k\|)\right)}{\sum_j \exp\left(\frac{1}{\tau}(\mathbf{f}_i/\|\mathbf{f}_i\|)^T(\mathbf{c}_j/\|\mathbf{c}_j\|)\right)}\,, \tag{2}$$

and τ denotes the temperature parameter in Softmax normalization. Obviously, $\mathbf{p}_i$ is measured by the cosine similarity between $\mathbf{f}_i$ and all prototype vectors. During training, the prototypical classifier h is trained by jointly optimizing a supervised loss on labeled samples and a self-supervised loss on all samples, i.e.,

$$\mathcal{L}_{cls}^s = \frac{1}{2|B^l|}\sum_{i\in B^l} l(\mathbf{y}_i, \mathbf{p}_i^a) + l(\mathbf{y}_i, \mathbf{p}_i^b) \tag{3}$$

and

$$\mathcal{L}_{cls}^u = \frac{1}{|B|}\sum_{i\in B} l(\hat{\mathbf{y}}_i^a, \mathbf{p}_i^b) - \epsilon H(\bar{\mathbf{p}}_i), \tag{4}$$

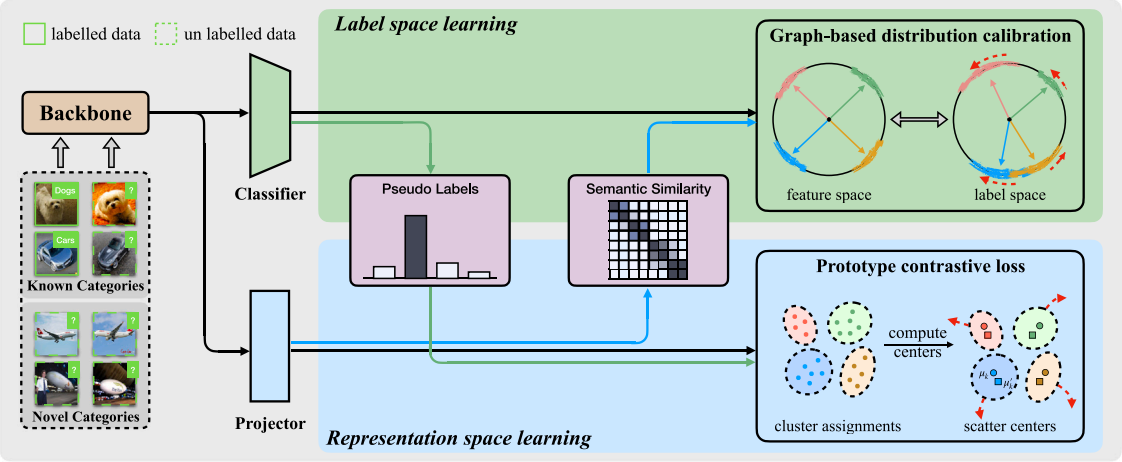

Fig. 2. The overarching framework of MSGCD involves joint learning of a representation space and label space, with enhancements derived from pseudo-label information and semantic similarity information, respectively.

where $l(\mathbf{y}_i, \mathbf{p}_i)$ denotes the cross-entropy loss, $\bar{\mathbf{p}}_i = \frac{1}{2}(\mathbf{p}_i^a + \mathbf{p}_i^b)$, and $H(\bar{\mathbf{p}})$ denotes the mean-entropy maximization regularizer [34]. Moreover, $\hat{\mathbf{y}}_i^a$ denotes the soft pseudo-labels produced by the view $\mathbf{x}_i^a$. The overall objective for label space learning is

$$\mathcal{L}_{cls} = \lambda \mathcal{L}_{cls}^s + (1-\lambda)\mathcal{L}_{cls}^u \tag{5}$$

where λ is a trade-off parameter. In particular, label prediction $\mathbf{p}$ directly reveals the classification results in the testing stage.

3.4. Representation space learning

Similarly, for representation space learning, we engage in both supervised contrastive learning using labeled samples and self-supervised contrastive learning across all training instances. Specifically, the supervised contrastive learning loss is computed as follows:

$$\mathcal{L}_{rep}^s = \frac{1}{|B^l|}\sum_{i\in B^l}\frac{1}{|\mathcal{N}_i|}\sum_{q\in\mathcal{N}_i} -\log\frac{\exp(\mathbf{z}_i^T\mathbf{z}_q/\tau_s)}{\sum_j^{i\neq j}\exp(\mathbf{z}_i^T\mathbf{z}_j/\tau_s)}. \tag{6}$$

where $\mathcal{N}_i$ denotes the positive pairs of $\mathbf{x}_i$. And, the self-supervised contrastive learning loss is

$$\mathcal{L}_{rep}^u = \frac{1}{|B|}\sum_{i\in B} -\log\frac{\exp(\mathbf{z}_i^T\mathbf{z}_i'/\tau_u)}{\sum_j^{i\neq j}\exp(\mathbf{z}_i^T\mathbf{z}_j'/\tau_u)}, \tag{7}$$

where τ_u, τ_s are two temperature parameters. The overall objective for representation space learning is

$$\mathcal{L}_{rep} = \lambda\mathcal{L}_{rep}^s + (1-\lambda)\mathcal{L}_{rep}^u. \tag{8}$$

Note. SimGCD [16] proposes to jointly learn the label space and representation space. And, the overall objective is

$$\mathcal{L}_{SimGCD} = \mathcal{L}_{cls} + \mathcal{L}_{rep}. \tag{9}$$

Although SimGCD is notably simple, its effectiveness is demonstrated by outperforming several state-of-the-art (SOTA) methods, as illustrated in Fig. 1. However, we contend that SimGCD does not fully exploit the potential of joint learning. In fact, the pseudo-labels generated from the label space can serve as weakly supervised information for representation space learning. Conversely, the semantic similarity information derived from the representation space proves valuable for label space learning.

3.5. Mutual-support mechanism

Our proposed mutual-support mechanism aims to regulate the outputs of the label space and representation space by leveraging valuable information from each other. This mechanism comprises two components: graph-based distribution calibration and pseudo-label-based prototype contrastive learning. On one hand, the high-quality pseudo-labels generated from the label space furnish valuable supervised information for learning discriminative representations. On the other hand, discriminative representations offer meaningful semantic similarities among samples, guiding the model to generate more reliable pseudo-labels.

3.5.1. Graph-based distribution calibration for label space learning

As previously discussed, the representation space can generate clustering-friendly representations that unveil meaningful semantic similarities among samples. Motivated by this observation, we introduce a graph-based distribution calibration for label space learning. Given the representations $\{\mathbf{z}_i\}_{i=1}^{2|B|}$ for the batch of samples, we construct a similarity graph by creating a similarity matrix $\mathbf{W}$ of size $2|B|\times 2|B|$. Subsequently, we employ the graph-based distribution calibration loss to regulate the output of the label space:

$$\mathcal{L}_{dc} = w_{ij}d_{ij}^2 + (1-w_{ij})(\delta - d_{ij})_+^2 \tag{10}$$

where w_{ij} is the semantic similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$, and d_{ij} is the euclidean distance between probability distribution vectors $\mathbf{p}_i$ and $\mathbf{p}_j$. The first term encourages samples with larger semantic similarity to have similar label predictions. The second term pushes the samples considered more dissimilar in the representation space more heavily out of the margin in the label space.

Indeed, the quality of estimated semantic similarities holds paramount importance in label space learning. However, estimating reliable semantic similarities becomes challenging due to the presence of numerous unlabeled samples. To address this challenge, we propose to utilize contextualized similarity to capture common patterns frequently observed within each class. Specifically, the contextualized semantic similarity of a pair of samples $(\mathbf{z}_i, \mathbf{z}_j)$ is a combination of their pairwise similarity w_{ij}^p and contextual similarity w_{ij}^c. Formally, the pairwise similarity is

$$w_{ij}^p = \exp\left(-\frac{\|\mathbf{z}_i^t - \mathbf{z}_j^t\|_2^2}{\sigma}\right), \tag{11}$$

where σ represents the Gaussian kernel bandwidth. Contextual similarity operates on the assumption that a pair of samples with larger overlapping contexts are more likely to be semantically similar. One approach to defining context is to consider the nearest neighbors of a sample. Nonetheless, the closest neighbors can be unreliable, particularly in cases where the representation space lacks discriminative attributes. To handle this issue, we instead adopt k-reciprocal nearest

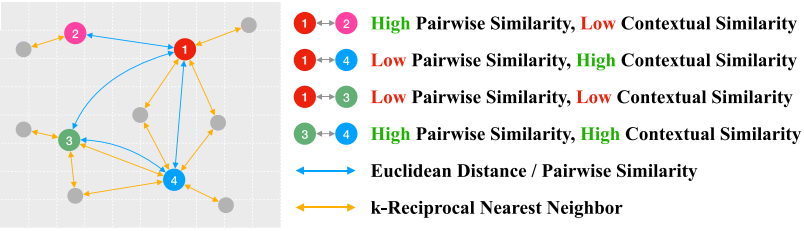

Fig. 3. How to determine the semantic similarity according to the relation between two samples.

neighbors to serve as the context [35–38]. The k-reciprocal nearest neighbors of a sample $\mathbf{x}_i$ in representation space is defined as:

$$R_k(\mathbf{z}_i) = \{\mathbf{z}_j | (\mathbf{z}_j \in N_k(\mathbf{z}_i)) \cap (\mathbf{z}_i \in N_k(\mathbf{z}_j))\}, \tag{12}$$

where $N_k(\mathbf{z}_i)$ is k nearest neighbors of $\mathbf{z}_i$. Further, we calculate the contextual similarity using asymmetric Jaccard similarity

$$\tilde{w}^c_{ij} = \frac{|R_k(\mathbf{z}_i) \cap R_k(\mathbf{z}_j)|}{|R_k(\mathbf{z}_i)|}, \forall \mathbf{z}_j \in R_k(\mathbf{z}_i). \tag{13}$$

In addition, we expect that similar samples exhibit similar spatial relationships. Hence, we adopt the idea of query expansion [38] to reformulate the contextual similarity as

$$\hat{w}^c_{ij} = \frac{1}{N_{k/2}(\mathbf{z}_i)} \sum_{h \in N_{k/2}(\mathbf{z}_i)} \tilde{w}^c_{hj}. \tag{14}$$

To guarantee the symmetry, the final contextual similarity w^c_{ij} is given by

$$w^c_{ij} = \frac{1}{2}(\hat{w}^c_{ij} + \hat{w}^c_{ji}). \tag{15}$$

We combine the pairwise similarity and contextual similarity together to obtain the final semantic similarity

$$w_{ij} = \frac{1}{2}(w^p_{ij} + w^c_{ij}). \tag{16}$$

The semantic similarity explores both the distance relationships and contextual relationships. For any pair of samples, as shown in Fig. 3, four main cases always exist, which are summarized as follows:

- *Case 1*: Despite a small Euclidean distance between ① and ②, and the absence of shared nearest neighbors, they exhibit low contextual similarity.
- *Case 2*: In contrast to *Case 1*, ① and ③ share many nearest neighbors but are far apart from each other. Consequently, they possess higher contextual similarity but lower pairwise similarity.
- *Case 3*: ① and ④ represent the worst case as they exhibit lower pairwise similarity, a large Euclidean distance, and no shared neighbors.
- *Case 4*: ③ and ④ share the k-reciprocal nearest neighbors and are close to each other with a small Euclidean distance. Thus, they exhibit both contextual and pairwise similarity.

3.5.2. Pseudo-labels-based prototype contrastive learning for representation space learning

As presented in Table 2, the predictions of all methods exhibit bias towards 'Old' categories. We posit that the primary reason for this bias lies in the absence of constraints over 'New' categories during training. Consequently, both the representation space and label space demonstrate small inter-category variances in 'New' categories, as depicted in Fig. 4. An alternative approach is to enforce the features from different categories to be as far apart as possible in the representation space. Furthermore, the label space can be indirectly enhanced by utilizing our proposed graph-based distribution calibration. To address this, we propose to regularize the representation space using pseudo-labels generated from the label space. Specifically, given a mini-batch of training data, we employ the classifier to assign pseudo-labels for unlabeled samples. These pseudo-labels are then combined with labeled samples to calculate two view-specific centers for each class:

$$\boldsymbol{\mu}^a_k = \frac{\sum_{\mathbf{z}^a_i \in C_k} \mathbf{z}^a_i}{\| \sum_{\mathbf{z}^a_i \in C_k} \mathbf{z}^a_i \|_2} \tag{17}$$

and

$$\boldsymbol{\mu}^b_k = \frac{\sum_{\mathbf{z}^b_i \in C_k} \mathbf{z}^b_i}{\| \sum_{\mathbf{z}^b_i \in C_k} \mathbf{z}^b_i \|_2}, \tag{18}$$

where C_k denotes the kth category. Then, we construct a prototype contrastive loss for these prototypes. That is,

$$\mathcal{L}_{pro} = \frac{1}{K} \sum_{i \in K} -\log \frac{\exp((\boldsymbol{\mu}^a_i)^T \boldsymbol{\mu}^b_i / \tau)}{\sum_j^{j \neq i} \exp((\boldsymbol{\mu}^a_i)^T \boldsymbol{\mu}^b_j / \tau)}. \tag{19}$$

Intuitively, the prototype contrastive loss encourages prototypes from the same category to move together, facilitating cohesion, while allowing prototypes from different categories to separate. In the realm of traditional representation space learning for GCD, the absence of pseudo-label information results in an inability to explore inter-class correlations, potentially leading to ambiguous representations. The pseudo-labels-based prototype contrastive loss addresses this limitation by minimizing inter-cluster similarity, promoting discriminative representations, and indirectly constraining the output of the label space.

3.5.3. Overall loss for mutual-support mechanism

The overall objective for our proposed *mutual-support mechanism* is

$$\mathcal{L}_{ms} = \alpha \mathcal{L}_{pro} + \beta \mathcal{L}_{dc}, \tag{20}$$

where α and β are two trade-off hyper-parameters and will be discussed in the next section. Besides, the Algorithm 1 shows the PyTorch-like pseudo-code of our MSGCD.

4. Experiments

4.1. Datasets and evaluation metric

We validate the effectiveness of MSGCD on four widely used datasets: CIFAR-100 [39], a standard image classification dataset, and CUB [40], Stanford Cars [41], FGVC-Aircraft [42], three more challenging fine-grained image classification datasets. All datasets are partitioned into labeled and unlabeled segments. Following the approach in [8], we designate a subset comprising half of the categories as the labeled categories (referred to as 'Old' categories) denoted as $\mathcal{Y}_l$. Half of the samples from these labeled class subsets are utilized to construct the labeled set $\mathcal{D}_l$, while the remaining samples form the unlabeled set $\mathcal{D}_u$. Detailed statistics and the division of datasets are summarized in Table 1.

For evaluation, we obtain the clustering accuracy (ACC) by comparing the predicted labels $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$, i.e.,

$$ACC = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{\mathbf{y}} = \delta(\mathbf{y})), \tag{21}$$

where N is the data scale and $\delta(\cdot)$ is the optimal mapping function that assigns the cluster results to ground truth.

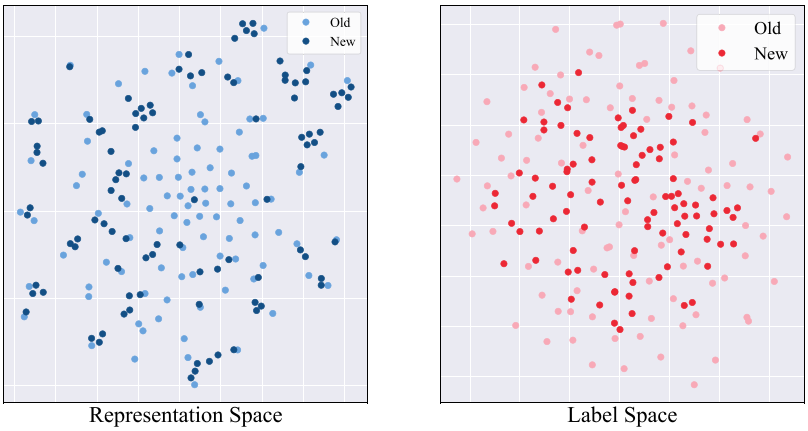

Fig. 4. t-SNE visualization of SimGCD on *Stanford Cars*. For representation space, each point stands for a category center, i.e., the mean of representations from the same class. For label space, each point stands for a parametric prototype.

Algorithm 1 Pseudo code for MSGCD

```
# f, g, h: backbone, projector and classifier
# all_z, Y_l: all features with N*d, known labels
# cat: matrix concatenation

for epoch in range(max_epoch):
  # training step
  for x, uq_idxs in loader:
    # random augmentations
    x1 = aug(x); x2 = aug(x)

    z1 = g(f(x1)); z2 = g(f(x2)) # projector
    p = h(f(x1)); q = h(f(x2)) # classifier

    # label space learning
    cls_loss = L_cls(p, q, Y_l)

    # representation space learning
    rep_loss = L_rep(z1, z2, Y_l)

    # mutual support mechanism
    ms_loss = L_ms(cat([z1, z2]), cat([p, q]))

    loss = cls_loss + rep_loss + ms_loss
    loss.backward()
```

4.2. Implementation details

Following previous works [8,13,16], we adopt the ViT-B-16 pre-trained by DINO [43] as a backbone network. The output [CLS] token with 768 dimensions serves as the feature representation for each input image. Fine-tuning is exclusively performed on the last transformer block for all methods. During the training phase, two views with random augmentations are fed to the model. Additionally, empirical findings indicate that a warm-up period for contrastive learning leads to improved performance. Therefore, mutual support is disabled in the first 60 epochs for all experiments. The best model is selected based on its performance on a validation set, formed using the test splits of each dataset. The performance on the unlabeled dataset is determined by selecting the best 'New' results. Mini-batches are constructed using nearest neighbors by computing the output of the projector and updating every epoch. Each mini-batch comprises 125 images, with 25 samples and their four nearest neighbors. We train the network for 200 epochs on each dataset with a cosine decay schedule with the initial learning rate of 0.1. For a fair comparison, aligning with [16], we set the trade-off factor λ as 0.35, temperature parameters τ_u, τ_s as 0.07, 1.0, respectively. The σ in Eq. (11) is set to 1 and $\tau = 0.5$ in Eq. (19). Other hyper-parameters (i.e. α, β) will be discussed later. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU.

Table 1
Statistics and separation of datasets for GCD.

Dataset	$\lvert\mathcal{Y}_l\rvert$	$\lvert\mathcal{D}_l\rvert$	$\lvert\mathcal{Y}_u\rvert$	$\lvert\mathcal{D}_u\rvert$
CIFAR-100	80	20 000	100	30 000
CUB	100	1500	200	4500
Stanford cars	98	2000	196	6100
FGVC-Aircraft	50	1700	100	5000

4.3. Comparison with state-of-the-art

In this test, we compare MSGCD to 9 representative GCD methods, including four parametric classifiers (RS+ [29], UNO+ [31], ORCA+ [9], PIM [10]), four non-parametric classifiers (GCD [8], DCCL [13], GPC [44] and PromptCAL [32]), and a joint-learning-based method (SimGCD [16]). Table 2 summarizes the experimental results on four benchmark datasets, where the best results are highlighted in bold. Moreover, the last row (Δ) of Table 2 shows the improvements of MSGCD over SimGCD. Upon reviewing Table 2, three conclusions can be drawn.

- Parametric classifiers often exhibit bias towards 'Old' categories, whereas non-parametric classifiers excel at recognizing 'New' categories. For instance, the performance differences between parametric classifiers and non-parametric classifiers are −7.3% (56.2% − 64.9%) and 11.3% (75.7% − 64.4%) on 'CUB-New' and 'CUB-Old', respectively. One of the main reasons for this discrepancy is that in pseudo-labels-based parametric classification methods, supervised information directly influences the parametric classifier, making it more inclined to assign samples to the 'Old' categories. Consequently, these methods exhibit performance improvements specifically on the 'Old' categories. In contrast, representation-based non-parametric classification methods prioritize representation learning, allowing them to more accurately capture the true distribution of the samples. Moreover, these methods are adept at learning a robust representation even for samples from unseen classes. Therefore, they have stable performances on 'New' categories.
- Methods based on joint learning can integrate the advantages of both parametric classifiers and non-parametric classifiers. For instance, both SimGCD and MSGCD demonstrate superior results in both 'New' and 'Old' categories. Notably, the advantages over other methods are particularly prominent in datasets such as Stanford Cars and FGVC-Aircraft.
- MSGCD consistently enhances SimGCD across different datasets, with maximum performance gains reaching 5.1% on 'Old' categories, and 4.9% on 'New' categories. The effectiveness of our

Table 2
Comparison results (%) with state-of-the-art methods. The best results are **bold**.

Type	Methods	Venue	Stanford cars			FGVC-Aircraft			CIFAR-100			CUB		
			All	Old	New	All	Old	New	All	Old	New	All	Old	New
Parametric	RS+	ICLR'20	28.3	61.8	12.1	26.9	36.4	22.2	58.2	77.6	19.3	33.3	51.6	24.2
	UNO+	ICCV'21	35.5	70.5	18.6	40.3	56.4	32.2	69.5	80.6	47.2	35.1	49.0	28.1
	ORCA	ICLR'22	23.5	50.1	10.7	22.0	31.8	17.1	69.0	77.4	52.0	35.3	45.6	30.2
	PIM	ICCV'23	43.1	66.9	31.6	–	–	–	78.3	84.2	66.5	62.7	**75.7**	56.2
Non-parametric	k-means	–	12.8	10.6	13.8	16.0	14.4	16.8	52.0	52.2	50.8	34.3	38.9	32.1
	GCD	CVPR'22	39.0	57.6	29.9	45.0	41.1	46.9	73.0	76.2	66.5	51.3	56.6	48.7
	DCCL	CVPR'23	43.1	55.7	36.2	–	–	–	75.3	76.8	70.2	63.5	60.8	**64.9**
	PromptCAL	CVPR'23	50.2	70.1	40.6	52.2	52.2	52.3	81.2	84.2	75.3	62.9	64.4	62.1
	GPC	ICCV'23	42.8	59.2	32.8	46.3	42.5	47.9	75.4	**84.6**	60.1	55.4	58.2	53.1
Joint-learning	SimGCD	ICCV'23	53.8	71.9	45.0	54.2	59.1	51.8	80.1	81.2	77.8	60.3	65.6	57.7
	MSGCD	–	**57.7**	**75.5**	**49.9**	**56.4**	**64.1**	**52.6**	**81.4**	82.5	**79.0**	**63.6**	70.7	60.0
	Δ		3.9↑	3.6↑	4.9↑	2.2↑	5.0↑	0.8↑	1.3↑	1.3↑	1.2↑	3.3↑	5.1↑	2.3↑

Table 3
Ablation study on the different components of our MSGCD.

	$\mathcal{L}_{pro}$	$\mathcal{L}_{dc}$	CUB			Stanford cars			FGVC-Aircraft		
			All	Old	New	All	Old	New	All	Old	New
(1)			60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8
(2)	✓		62.2	69.9	58.5	56.4	**76.2**	46.9	55.2	63.8	50.9
(3)		✓	61.8	69.7	57.9	55.9	75.1	46.6	55.6	63.2	51.8
(4)	✓	✓	**63.6**	**70.7**	**60.0**	**57.7**	75.5	**49.9**	**56.4**	**64.1**	**52.6**

proposed mutual-support mechanism is evident in improving joint-learning-based methods. This observation underscores that despite SimGCD combining the classifier and projector, these two components essentially operate in parallel with the backbone. Consequently, the inconsistency in distributions between the representation space and label space compromises the performance of the method. In contrast, our proposed mutual-support mechanism facilitates communication between these two distinct modules, enabling the synergistic utilization of their respective strengths and ultimately enhancing the algorithm's performance.

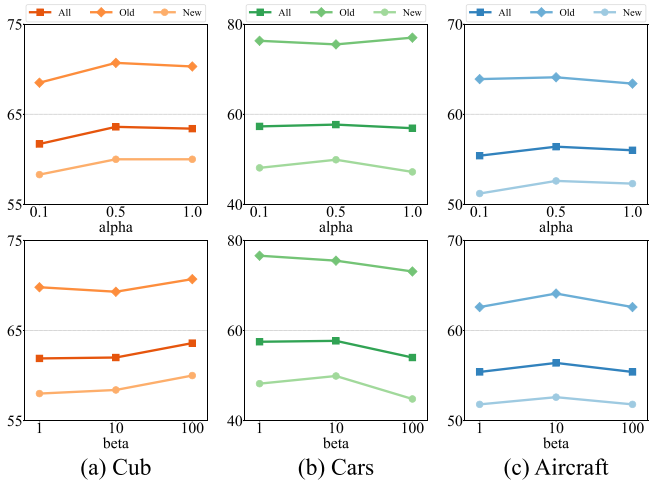

Fig. 5. Impact of the loss weights on 'All', 'Old', and 'New' categories.

4.4. Ablation study

Table 3 verifies the key components of MSGCD and shows their effectiveness. It should be noted that experiment (1) was the same as SimGCD, which serves as the baseline in our ablation experiments.

4.4.1. Effectiveness of pseudo-labels-based prototype contrastive learning

The comparison between (1) and (2) provides strong evidence supporting the effectiveness of utilizing pseudo-labels to guide the representation space. The analysis demonstrates the crucial role of pseudo-labels in improving the classification performance of the 'Old' categories. This improvement is evident across three datasets, with respective gains of 4.3%, 5.3%, and 4.2%. However, when it comes to the 'New' categories, the observed performance gains are relatively modest at 0.8% and 1.9%, respectively. The primary reason behind this discrepancy lies in the classifier's limited ability to effectively distinguish the 'New' categories and generate meaningful pseudo-labels for representation learning. As a result, the impact of pseudo-labels on enhancing the performance in the 'New' categories is not as pronounced as in the 'Old' categories.

4.4.2. Effectiveness of graph-based distribution calibration

The comparison between (1) and (3) provides substantial evidence supporting the effectiveness of employing graph-based distribution calibration to guide the label space. This module enhances the classification precision of the 'New' categories while maintaining the performance of the 'Old' categories. This outcome is logical and aligns with the earlier discussion, as clustering-friendly representations prove valuable in accurately categorizing the 'New' categories. Additionally, it should be noted that during the training process, the supervised information exclusively pertains to the Old class. Consequently, the embeddings in the representation space for the 'Old' class data inherently exhibit a high level of precision.

4.4.3. Effectiveness of mutual-support mechanism

The comparison between (4) and (1), (2), (3) provides strong validation for the effectiveness of our proposed mutual-support mechanism. MSGCD consistently enhances the performance of the baseline model in both the 'New' and 'Old' categories. Moreover, the observed performance gains surpass those achieved by solely relying on either pseudo-labels-based prototype contrastive learning or graph-based distribution calibration.

Through analysis, we uncover that the utilization of *mutual-support*, as opposed to *unidirectional support*, is crucial. This finding highlights the significance of bidirectional interaction and collaboration between the classifier and projector modules within MSGCD. The *mutual-support mechanism* allows for the comprehensive integration of their respective strengths, leading to superior performance improvements across both the 'New' and 'Old' categories.

4.5. Hyper-parameter analysis

In this section, we analyze the impact of the hyper-parameters in MSGCD, including loss trade-off parameters, α, β, k nearest neighbors hyper-parameters used in Eq. (10) and temperature factor τ in Eq. (19).

Table 4
Impact of nearest neighbors k in Eq. (10).

TOP-K	Stanford cars			FGVC-Aircraft			CUB		
	All	Old	New	All	Old	New	All	Old	New
5	56.9	**78.0**	46.7	**56.7**	63.6	**53.3**	62.0	**71.0**	57.5
10	57.0	74.7	48.4	55.9	63.4	52.2	63.2	70.7	59.5
15	**57.7**	75.5	**49.9**	56.4	**64.1**	52.6	**63.4**	69.9	**60.2**
20	50.2	60.2	45.2	55.8	62.7	52.3	63.3	70.7	60.0

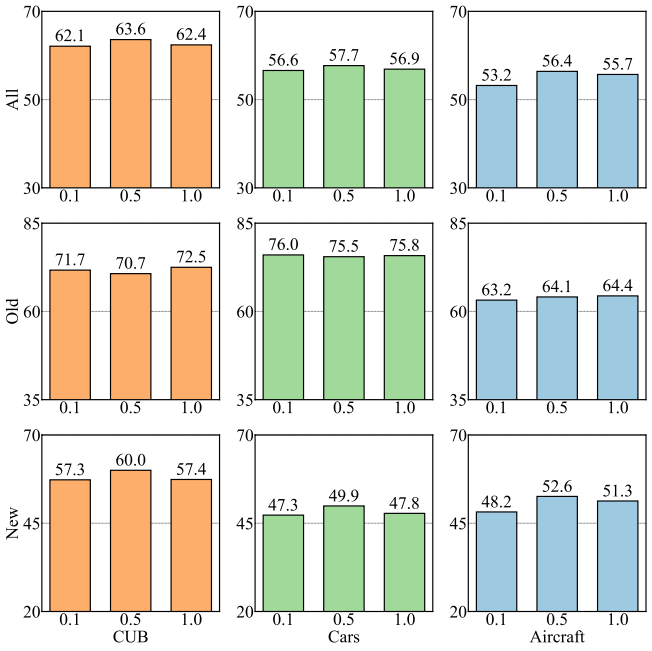

Fig. 6. Impact of temperature factor of τ on three datasets.

4.5.1. Effect of the trade-off parameters α and β

In this experiment, we investigate the impact of varying the values of α and β in $S_\alpha = \{0.1, 0.5, 1.0\}$ and $S_\beta = \{1, 10, 100\}$, respectively, corresponding to their values during the training phase. The results are illustrated in Fig. 5, where we observe a relatively small performance gap between different parameter settings. On one hand, when α and β are assigned small values, the interaction between the classifier and projector is suppressed, leading to the degeneration of MSGCD into SimGCD. This occurs because tiny values restrict the *mutual taught* between the two modules. In extreme cases, such as when $\alpha = \beta = 0$, MSGCD completely degrades into SimGCD. On the other hand, selecting large trade-off parameters can also have a detrimental effect on the model's performance. This is due to the potential noise introduced by the pseudo-supervised information used in $\mathcal{L}_{ms}$, as well as the possibility of constructing misleading semantic similarities from the representation space. Overall, we find that MSGCD exhibits robustness to parameter selection within appropriate intervals. In our experiments, we set $\alpha = 0.5$ and $\beta = 10$ as fixed values for all conducted experiments, as they have demonstrated favorable performance.

4.5.2. Effect of nearest neighbors k in Eq. (10)

The investigation on the impact of nearest neighbors, denoted by the parameter k, is presented in Table 4. Upon reviewing the results in Table 4, it becomes evident that MSGCD consistently achieves high performance for k values below 15. Moreover, MSGCD outperforms all other GCD methods when k is less than 15. This observation aligns with our intuitive understanding. When k exceeds 15, each sample considers an excessive number of neighbors. As a result, connections are formed between points that do not belong to the same class, leading to the introduction of noise in the semantic similarity.

In sum, a smaller k ignores the structural information between samples, thereby reducing the ability to mine latent semantic information. On the other hand, a larger k value will include too many redundant samples, resulting in consideration of false positive samples that do not belong to the same categories. Consequently, this noise adversely affects the classification accuracy of the impression classifier.

4.5.3. Effect of temperature factor of τ

To investigate the influence of temperature factor τ on model performance, we vary it in the set $S_\tau = \{0.1, 0.5, 1.0\}$. As shown in Fig. 6, our model achieves the best performance under different settings on all datasets and evaluation metrics, which can robustness of center scatter loss. Based on our results, we finally suggest using a default value of $\tau = 0.5$ in other all experiments.

4.6. In-depth analysis

4.6.1. Analysis on contextualized semantic similarity

To gain a deeper understanding and analyze the contextualized semantic similarity, we present it in Fig. 7. The first column represents the anchors, while the remaining columns display the top-7 image pairs corresponding to semantic similarity. As we can each, the top-5 nearest neighbors belong to the same categories, with similarities all exceeding 0.7. Regarding the last two columns, they do not share the same labels as the anchors, and their similarities are significantly lower than those of the 5th nearest neighbors. This demonstrates the effectiveness of contextualized semantic similarity. In other words, even though we know that the last two columns belong to different categories, they are still visually difficult to distinguish. Therefore, we can explore the valuable information in the representation space and enhance the quality of embeddings in the label space.

4.6.2. Effect of mutual-support mechanism

In this subsection, to briefly illustrate how the *mutual-support mechanism* works, we use TSNE to visualize the category centers in the representation space under the different models (SimGCD and MSGCD). As shown in left side of Fig. 8, we find that the inter-class distance with respect to 'New' category centers (dark blue) is small in SimGCD, and the distribution of 'All' centers in MSGCD is more uniform than SimGCD. The above compact centers distribution may be the main reasons limiting the performance. Therefore, an intuitive idea is that makes the cluster centers far from each other. Since it can enlarge the inter-class distance and obtain a better representations, enhancing the effectiveness of discovering unseen categories.

In order to further analyze the advantages of MSGCD over SimGCD, we provide another visualization experimental result on *Stanford Cars*. Specifically, we estimate the probability density function of pairwise distances between prototypical classifiers and then use kernel density estimation (KDE) [45] to smooth it. We report the results of the supervised model (Supervised) and initial pre-trained model (Init) for reference. Upon observing Fig. 9, we have made several interesting observations. For the initial pre-trained model, the distribution of distances between each prototype is random. In contrast, for the supervised model, the distribution of distances is highly concentrated, indicating that the pairwise distances between any two prototypes are almost identical. This observation directly indicates that the prototypes are uniformly distributed in the label space, which can be considered an ideal distribution. Conversely, when considering SimGCD and MSGCD, we can observe that MSGCD exhibits a closer alignment with the ideal distribution, suggesting that the prototypes in MSGCD are distributed more uniformly within the label space. Consequently, a significant factor contributing to the superior performance of MSGCD over SimGCD is

Fig. 7. Top-7 image pairs sorted by their contextualized semantic similarity on (a) FGVC-Aircraft and (b) CUB. For each images with blue boundary are anchors, green boundary are of the same class as anchors and those with red boundary are of a different class from anchors.

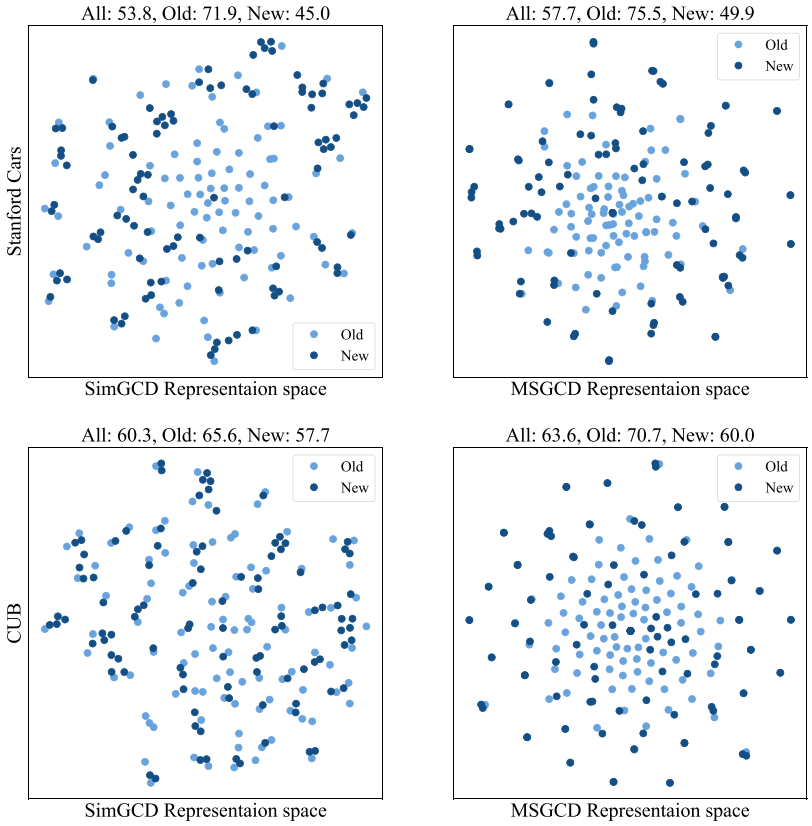

Fig. 8. TSNE visualization for category centers in the representation spaces of SimGCD and MSGCD.

the utilization of the *mutual-support mechanism*, which promotes a more even distribution of prototypes within the label space.

4.6.3. Representation space v.s. label spaces

In this test, we compare the results between classifier and projector. For projector in representation space, we obtain the results by adopting semi-supervised k-means++. For classifier, we directly leverage its outputs as clustering assignments. According to Table 5, we summarize the following conclusions:

- Joint-learning-based methods (SimGCD and MSGCD) outperform than non-parametric method (GCD) both on two modules, since they can combine the advantages of label and representation space learning.

Table 5
Comparison result (%) between *Classifier* and *Projector*. The best results are **bold**.

Method	Output	Stanford cars			FGVC-Aircraft			CIFAR-100			CUB		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD	*Projector*	39.0	57.6	29.9	45.0	41.1	46.9	73.0	76.2	66.5	51.3	56.6	48.7
SimGCD	*Projector*	48.3	65.2	40.0	52.8	56.3	51.0	73.1	78.6	62.1	56.2	60.4	54.2
	Classifier	53.8	71.9	45.0	54.2	59.1	51.8	80.1	81.2	77.8	60.3	65.6	57.7
MSGCD	*Projector*	56.7	71.6	49.5	53.3	61.8	48.8	80.4	82.3	76.4	58.6	66.4	54.7
	Classifier	**57.7**	**75.5**	**49.9**	**56.4**	**64.1**	**52.6**	**81.4**	**82.5**	**79.0**	**63.6**	**70.7**	**60.0**

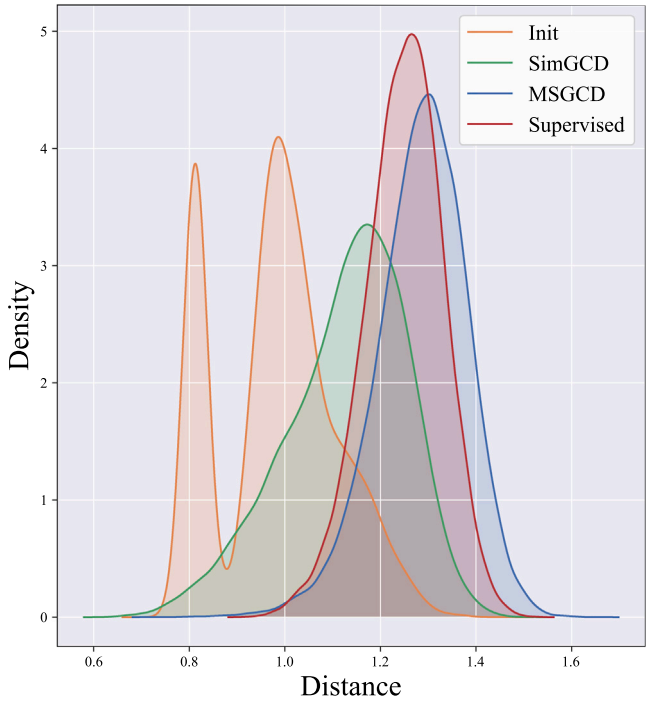

Fig. 9. Kernel density estimation of pairwise distance distribution of the prototypical classifier in label space on *Stanford Cars*.

- MSGCD outperforms than SimGCD no matter the results are obtained from *projector* or *classifier*, which denotes that the mutual-support mechanism is effective for improving joint-learning-based methods.
- Classifier are feasible for predicting 'Old' categories yet the non-parametric method tends to predict 'New' categories. For example, MSGCD's *projector* achieves the 9.5% and 14.3% improvements on Cars and CIFAR-100. In contrast, the best results of MSGCD's *classifier* focuses on 'Old' categories. There exists 3.6% and 4.3% performance gains than the second best results. This once again confirms the distinct roles of the representation space and label space, where the former is capable of learning meaningful representations while the latter is more sensitive to supervised information as mentioned above.

4.6.4. Complete performance of MSGCD for different training stage

Fig. 10 illustrates the performance of our model during the training process on 'All', 'Old', and 'New' data. Our performance stabilizes around epoch = 60 and exhibits improvements thereafter, particularly after epoch = 70. Empirically, we observe that SimGCD, employed as the baseline, demonstrates stable performance after 60 epochs. Consequently, we introduced the mutual support mechanism at epoch = 60. After the introduction of the mutual support mechanism, fluctuations in both the 'Old' and 'New' categories are observed in MSGCD. This occurs because the mechanism reorganizes the sample distribution in both the representation space and label space, thereby establishing cross-space consistency. MSGCD allows the model to overcome local optima and attain superior results. Finally, when the positions of the cluster centers are relatively stable, MSGCD can achieve optimal performance.

5. Conclusion

In this work, we introduced a novel framework, MSGCD, for generalized category discovery. MSGCD offers a novel perspective on integrating the strengths of parametric classification methods and non-parametric classification methods. Through experiments on four widely used datasets, MSGCD consistently outperforms the baseline by substantial margins, establishing state-of-the-art performance.

It should be noted that although the interactive method in this section can achieve good results, it is an unavoidable shortcoming that label space and representation learning use different network parameters. Therefore, how to integrate label learning and representation learning from the network structure is a challenging problem in the next future. Additionally, we aim to generalize MSGCD to tackle a more realistic setting where the number of new categories is unknown. Moreover, how to extend the idea of mutual-support mechanism to open-world image recognition, image segmentation, and even text and sequence data is one of the important challenges in the future works.

CRediT authorship contribution statement

Yu Duan: Writing – original draft, Methodology, Conceptualization. **Zhanxuan Hu:** Writing – original draft, Formal analysis, Conceptualization. **Rong Wang:** Writing – original draft, Supervision, Funding acquisition. **Zhensheng Sun:** Writing – review & editing, Supervision. **Feiping Nie:** Writing – review & editing, Supervision, Funding acquisition. **Xuelong Li:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0101902, in part by the Natural Science Basic Research Program of Shaanxi, China (Program No. 2021JM-071), in part by the National Natural Science Foundation of China under Grant 62176212, Grant 61936014 and Grant 61772427, Grant 62176203, and in part by the Fundamental Research Funds for the Central Universities, China under Grant G2019KY0501. This work is also supported by Zhijian Laboratory (Rocket Force University of Engineering), China.

Data availability

Data will be made available on request.

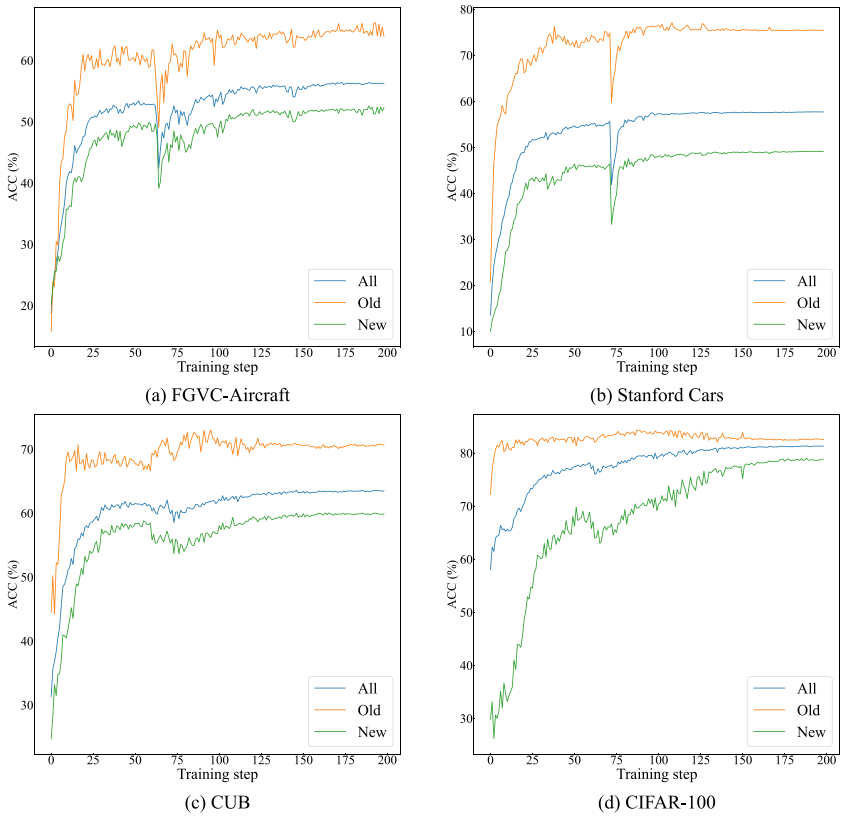

Fig. 10. Complete performance of MSGCD on four datasets for different training stage.

References

[1] X. Luo, W. Ju, M. Qu, Y. Gu, C. Chen, M. Deng, X.-S. Hua, M. Zhang, Clear: Cluster-enhanced contrast for self-supervised graph representation learning, IEEE Trans. Neural Networks Learn. Syst. 35 (1) (2022) 899–912.

[2] X. Li, Y. Fan, G. Lv, H. Ma, Area-based correlation and non-local attention network for stereo matching, Vis. Comput. 38 (11) (2022) 3881–3895.

[3] Q. Zhou, Q. Wang, Q. Gao, M. Yang, X. Gao, Unsupervised discriminative feature selection via contrastive graph learning, IEEE Trans. Image Process. (2024).

[4] E. Fini, P. Astolfi, K. Alahari, X. Alameda-Pineda, J. Mairal, M. Nabi, E. Ricci, Semi-supervised learning made simple with self-supervised clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3187–3197.

[5] X. Luo, Y. Zhao, Y. Qin, W. Ju, M. Zhang, Towards semi-supervised universal graph classification, IEEE Trans. Knowl. Data Eng. 36 (1) (2023) 416–428.

[6] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C.A. Raffel, E.D. Cubuk, A. Kurakin, C.-L. Li, Fixmatch: Simplifying semi-supervised learning with consistency and confidence, Adv. Neural Inf. Process. Syst. 33 (2020) 596–608.

[7] W. Ju, Y. Qin, Z. Qiao, X. Luo, Y. Wang, Y. Fu, M. Zhang, Kernel-based substructure exploration for next POI recommendation, in: 2022 IEEE International Conference on Data Mining, ICDM, IEEE, 2022, pp. 221–230.

[8] S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Generalized category discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7492–7501.

[9] K. Cao, M. Brbic, J. Leskovec, Open-world semi-supervised learning, in: International Conference on Learning Representations, 2021.

[10] F. Chiaroni, J. Dolz, Z.I. Masud, A. Mitiche, I. Ben Ayed, Parametric information maximization for generalized category discovery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1729–1739.

[11] M.N. Rizve, N. Kardan, M. Shah, Towards realistic semi-supervised learning, in: European Conference Computer Vision, Springer, 2022, pp. 437–455.

[12] S. Zhang, S. Khan, Z. Shen, M. Naseer, G. Chen, F.S. Khan, Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3479–3488.

[13] N. Pu, Z. Zhong, N. Sebe, Dynamic conceptional contrastive learning for generalized category discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7579–7588.

[14] B. Zhao, X. Wen, K. Han, Learning semi-supervised Gaussian mixture models for generalized category discovery, 2023, arXiv preprint arXiv:2305.06144.

[15] S. Hao, K. Han, K.-Y.K. Wong, CiPR: An efficient framework with cross-instance positive relations for generalized category discovery, 2023, arXiv preprint arXiv:2304.06928.

[16] X. Wen, B. Zhao, X. Qi, Parametric classification for generalized category discovery: A baseline study, in: IEEE International Conference on Computer Vision, 2023.

[17] W. Ju, X. Luo, M. Qu, Y. Wang, C. Chen, M. Deng, X.-S. Hua, M. Zhang, TGNN: A joint semi-supervised framework for graph-level classification, 2023, arXiv preprint arXiv:2304.11688.

[18] W. Xia, T. Wang, Q. Gao, M. Yang, X. Gao, Graph embedding contrastive multi-modal representation learning for clustering, IEEE Trans. Image Process. 32 (2023) 1170–1183.

[19] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C.A. Raffel, Mixmatch: A holistic approach to semi-supervised learning, Adv. Neural Inf. Process. Syst. 32 (2019).

[20] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, 2017, arXiv preprint arXiv:1710.09412.

[21] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj, et al., Freematch: Self-adaptive thresholding for semi-supervised learning, 2022, arXiv preprint arXiv:2205.07246.

[22] J. Li, C. Xiong, S.C. Hoi, Comatch: Semi-supervised learning with contrastive graph regularization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9475–9484.

[23] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, C. Xu, Simmatch: Semi-supervised learning with similarity matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14471–14481.

[24] K. Han, A. Vedaldi, A. Zisserman, Learning to discover novel visual categories via deep transfer clustering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8401–8409.

[25] Y.-C. Hsu, Z. Lv, Z. Kira, Learning to cluster in order to transfer across domains and tasks, 2017, arXiv preprint arXiv:1711.10125.

[26] Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, Z. Kira, Multi-class classification without multi-class labels, 2019, arXiv preprint arXiv:1901.00544.

[27] S. Lin, C. Liu, P. Zhou, Z.-Y. Hu, S. Wang, R. Zhao, Y. Zheng, L. Lin, E. Xing, X. Liang, Prototypical graph contrastive learning, IEEE Trans. Neural Networks Learn. Syst. (2022).

[28] X. Yang, X. Hu, S. Zhou, X. Liu, E. Zhu, Interpolation-based contrastive learning for few-label semi-supervised learning, IEEE Trans. Neural Networks Learn. Syst. (2022).

[29] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, A. Zisserman, Autonovel: Automatically discovering and learning novel visual categories, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 6767–6781.
[30] Z. Zhong, E. Fini, S. Roy, Z. Luo, E. Ricci, N. Sebe, Neighborhood contrastive learning for novel class discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10867–10875.
[31] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, E. Ricci, A unified objective for novel class discovery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9284–9292.
[32] S. Zhang, S. Khan, Z. Shen, M. Naseer, G. Chen, F.S. Khan, Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3479–3488.
[33] Y. Duan, J. He, R. Zhang, R. Wang, X. Li, F. Nie, Prediction consistency regularization for generalized category discovery, Inf. Fusion 112 (2024) 102547.
[34] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, N. Ballas, Masked siamese networks for label-efficient learning, in: European Conference on Computer Vision, Springer, 2022, pp. 456–473.
[35] H. Jegou, H. Harzallah, C. Schmid, A contextual dissimilarity measure for accurate and efficient image search, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[36] D. Qin, S. Gammeter, L. Bossard, T. Quack, L. Van Gool, Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors, in: CVPR 2011, IEEE, 2011, pp. 777–784.
[37] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.
[38] S. Kim, D. Kim, M. Cho, S. Kwak, Self-taught metric learning without labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7431–7441.
[39] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, 2009.
[40] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-UCSD birds 200, 2010.
[41] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 554–561.
[42] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, 2013, arXiv preprint arXiv:1306.5151.
[43] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.
[44] B. Zhao, X. Wen, K. Han, Learning semi-supervised Gaussian mixture models for generalized category discovery, 2023, arXiv preprint arXiv:2305.06144.
[45] G.R. Terrell, D.W. Scott, Variable kernel density estimation, Ann. Statist. (1992) 1236–1265.