AMEND: Adaptive Margin and Expanded Neighborhood for Efficient Generalized Category Discovery

Anwesha Banerjee, Liyana Sahir Kallooriyakath, Soma Biswas
Indian Institute of Science, Bengaluru
anweshabaner@iisc.ac.in, liyanasahir@iisc.ac.in, somabiswas@iisc.ac.in

Abstract

Generalized Category Discovery aims to discover and cluster images from previously unseen classes, in addition to classifying images from seen classes correctly. In this work, we propose a simple, yet effective framework for this task, which not only performs on-par or better with the current approaches but is also significantly more efficient in terms of computational requirements. Our first contribution is to use expanded neighborhood information in contrastive learning to generate robust and generalizable features. To generate more discriminative feature representations, especially for fine-grained datasets and confusing classes, we propose a class-wise adaptive margin regularizer that aims at increasing the angular separation among the prototypes of all classes. Extensive experiments on three generic as well as four fine-grained benchmark datasets show the usefulness of the proposed Adaptive Margin and Expanded Neighborhood (AMEND) framework.

1. Introduction

Due to the exponential increase in data, it is often not feasible to obtain precise annotations for vision-related tasks. However, in a practical scenario, some amount of labeled and unlabeled data may be available, consisting of images belonging to previously seen classes, as well as classes that are somewhat related but previously unseen. For instance, an image recognition system for self-driving cars may be trained to recognize a hundred classes of objects commonly observed on roads. However, it may also encounter similar objects from outside the classes it has been trained on, such as a new car brand or a new street sign. This is exactly what Generalized Category Discovery (GCD) aims to solve, i.e. discover and cluster images from previously unseen classes, in addition to correctly classifying images from the already seen classes. GCD [30] is a more challenging and realistic extension of the Novel Class Discovery (NCD) problem [11], where the goal is to discover and categorize the data present in the unlabeled data, which consists of data only from previously unseen classes. Though several approaches have been proposed in the literature [8, 24, 30, 36], we observe that there is a tradeoff between computational requirement and performance. Some of the approaches are quite efficient [30], but their performance is not at par with the the state-of-the-art prompting based approach [36], which is quite computationally intensive.

In this work, we propose a novel framework, termed AMEND (**A**daptive **M**argin and **E**xpanded **N**eighbourhoo**d**) for addressing the challenging GCD task. The proposed approach bridges this gap, thus achieving on-par or better performance with the state-of-the-art approaches, while being computationally efficient. Motivated by several recent approaches [39] [7] [34], [36], AMEND utilizes expanded neighbors of each instance along with its direct neighborhood information. This helps to compute robust feature representations in a contrastive learning framework, which leads to better clustering. The majority of the existing approaches [8, 26, 30, 33] treat all classes identically. However, for fine-grained datasets and for confusing classes, feature representations can be close to each other in feature space, thereby adversely affecting the clustering quality. This causes a decrease in final performance due to an increase in confusion. Our second contribution is a class-wise adaptive margin regularizer that aims to better separate the prototypes which act as representatives of each class. During training, we propose to increase the angular separation between the corresponding prototypes based on whether the prototypes are coming close to each other, thus ensuring less confusion between instances belonging to those classes. In summary, the contributions of this work are:

1. We propose a novel framework termed AMEND for the challenging GCD task.

2. We propose to incorporate expanded neighborhood information in the standard contrastive loss for unlabeled data to ensure better clustering of same class samples.

3. We also propose a class-adaptive margin regularizer to reduce the confusion between adjacent classes, thus improving the class discriminability and classification performance.
4. Extensive experiments on several generic and fine-grained benchmark datasets and comparisons with state-of-the-art techniques justify the effectiveness of the proposed framework.
5. The proposed AMEND framework is significantly less resource intensive compared to the existing state-of-the-art approaches.

In the following sections, we discuss the related work, followed by a description of the proposed framework along with results of the extensive evaluation and analysis.

2. Related Work

Here, we briefly discuss the related work in literature.

Category Discovery: Novel Category Discovery (NCD) [11] refers to the weakly supervised setting in which a labeled set of known classes and an unlabeled set of unknown classes are given during training, and we are tasked with discovering and classifying the instances of the unknown or 'novel' classes. Initial works in NCD focused on a two-stage approach [11, 13, 14], where the first stage focused on representation learning with just the labeled data, and the second stage involved transfer learning on the unlabeled data. More recent works [9, 10, 37, 39, 40] focus on learning representations for both the labeled and unlabeled data simultaneously with separate classification heads and objectives, specifically classification on the labeled data and clustering or classification with pseudo-labels on the unlabeled data. Notably, UNO [9] uses a unified loss function with the help of pseudo-labels using the Sinkhorn-Knopp [4] algorithm, while RankStats [10, 37] uses ranking-statistics to obtain pseudo-labels for the classification heads.

Generalised Category Discovery (GCD) was formalized by GCD [30] and ORCA [31], as a practical extension of NCD, where the unlabeled data also contains instances of classes that are seen in the labeled data, in addition to instances from completely novel classes. The baselines for NCD such as UNO [9] and RankStats [10] were adopted to the GCD setting by [30] by extending their classification heads. In [30], the unsupervised and supervised contrastive losses [3] are used to finetune a ViT [18] model pre-trained with DINO [2], which increases the similarity between the feature representations for different views of the same image, and between different images of the same class.

Recently, a few extensions of GCD have been proposed. XCon [8] first partitions the dataset into subsets by clustering them using kmeans, and then performs contrastive learning separately on each of the clusters now obtained, while OpenCon [26] adds an additional novel class loss by performing out-of-distribution detection to identify the instances belonging to novel classes with the help of prototype vectors. SimGCD [33] simplifies the framework presented by GCD [30] by reintroducing parametric classification with the help of prototype vectors and by applying self-distillation to obtain pseudo-labels.

Prompt tuning has emerged as a powerful technique in the field of Natural Language Processing (NLP) and has been extended to images with visual prompt learning (VPT) [16]. VPT involves fine-tuning embedded visual prompts using a pre-trained Vision Transformer (ViT) backbone supervised by downstream objectives, aiming to improve transfer learning performance. PromptCAL [36] was the first to adapt VPT to the GCD setting. It utilizes prompts to provide a weaker semantic supervision signal. In [24], conceptional contrastive learning is used, considering the relationships between instances and improving clustering accuracy compared to methods solely relying on instance-level contrastive learning.

Contrastive Learning: Contrastive Learning is a commonly used unsupervised learning technique which involves learning representations of data by contrasting similar and dissimilar pairs of data. CPC [23] introduced InfoNCE, a contrastive loss that can be used to learn representations of speech that is invariant to certain factors of variation, such as the identity of the speaker or the language of the speech. SimCLR [3] established the importance of data augmentations, a large batch size, and a modified loss function that encouraged the representations of two different views of the same image to be similar, while pushing apart the representations of different images. Weakly-supervised and completely supervised variations of this loss [17, 38] have also been proposed, which consider images of the same class to be positives if the labels are available, on top of using just the augmentations.

Neighborhood Clustering: Several works [34, 35] in Source-Free Domain Adaptation exploit the intrinsic neighborhood structure to learn better representations in the feature space. In the context of contrastive learning, [7, 25] use of the nearest neighbors of a sample in the feature space as positives to cover for more semantically meaningful variations within a class than just augmentations. In FNC [15], the authors show how performance can be improved significantly by identifying and repelling nearest neighbors that are false negatives and true positives appropriately. Notably, the NNCLR [7] loss uses the nearest neighbors of a sample stored in a support set over the training period as positives in the contrastive loss, which improves the performance significantly.

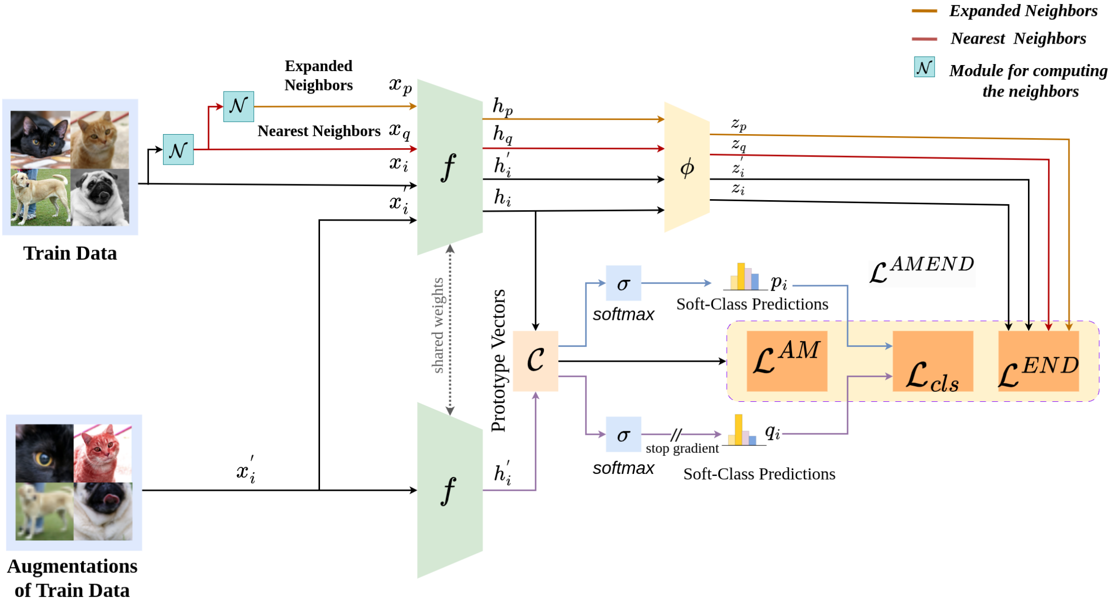

Figure 1. Illustration of the proposed AMEND framework. Images in a batch are passed through the ViT-DINO backbone (f) and the MLP projection head (ϕ) to obtain the corresponding feature representations $\mathbf{h}$ and $\mathbf{z}$ respectively. Here, $\mathcal{N}$ represents the module used for computing neighbors. An image $\mathbf{x}_i$ is initially subjected to module $\mathcal{N}$ to enumerate the nearest neighbors of the image. After obtaining the nearest neighbors $\mathbf{x}_q$, we again employ the module $\mathcal{N}$ for obtaining the expanded neighbors $\mathbf{x}_p$ (neighbors of the neighbors). The feature representations $(\mathbf{z}_i, \mathbf{z}_i', \mathbf{z}_q, \mathbf{z}_p)$ obtained from the MLP projection head (ϕ) are then used to compute the representation loss $\mathcal{L}^{END}$. The feature representation of the image and its augmentation ($\mathbf{h}_i$ and $\mathbf{h}_i'$) obtained from the ViT-DINO backbone are used to obtain soft class predictions with the help of Prototype Vectors. The Soft class predictions are then in turn used to compute the classifier loss $\mathcal{L}_{cls}$. Lastly,the adaptive margin regularizer, $\mathcal{L}^{\text{AM}}$ is calculated using the prototype vectors, $\mathbf{C}$.

Margin Regularization: Computing class-specific or adaptive margins has proved to be very successful in handling imbalances in training data. The work in [5] proposed a loss that uses an angular margin penalty which can be adjusted dynamically during training based on the angular distance between the feature vectors of the query and the positive samples to improve the discriminative power of the learned representations. In [12], the authors propose a max-margin framework based on an affinity measure in the Euclidean space that jointly reduces intra-class variations and maximizes inter-class distances . Extending this idea, [6] computes the regularizer from the class-wise training data distribution in the zero-shot sketch-based image retrieval setting in order to handle class imbalances by enforcing a broader margin for the classes with a lesser number of training samples. Our Adaptive Margin loss is an extension of the ideas presented in [6, 12] to the GCD setting, where we focus on spreading out the prototypes in the feature space taking the similarities between the prototypes into consideration.

3. Proposed AMEND Framework

Consider a dataset D containing two parts: a labeled set $\mathcal{D}_\mathcal{L}$ and an unlabeled set $\mathcal{D}_\mathcal{U}$, where the objective is to cluster all the images in the unlabeled set. The labeled and unlabeled set can be described as follows: $\mathcal{D}_\mathcal{L} = (\mathbf{x}_i, y_i)_{i=1}^{N} \in X \times Y_L$ and $\mathcal{D}_\mathcal{U} = \{\mathbf{x}_i\}_{i=1}^{M}$. The images in the unlabeled set belong to classes Y_U, and $Y_L \subset Y_U$. During training, the model does not have access to the labels in $\mathcal{D}_\mathcal{U}$, because it is tasked with predicting them at test time. Many techniques have been proposed [9, 11, 30] to calculate the total number of classes k in $\mathcal{D}_\mathcal{U}$. In this work, as in [36], we assume that k is known apriori or has been estimated using one of the existing approaches.

This proposed framework (AMEND) (Figure 1) makes two significant contributions, namely (i) Neighbors and expanded neighborhood for obtaining robust representations; and (ii) Adaptive Margin between the class prototypes to reduce the confusion between the adjacent classes, thereby improving the class discovery. We describe these two components in detail below:

3.1. Neighbors and Expanded neighborhood

Contrastive learning has been very successful in clustering input data, thereby discovering new classes for GCD task [8, 26, 30, 33]. For unlabeled data, usually, the original image and its augmentations are used as positive pairs, while for the labeled data (as in supervised contrastive learning), other samples from the same class can be utilized. Because of this difference in the way of handling samples from old (seen) and new (unseen) classes, the performance of the old classes is in general much better than that of the new classes. To bridge this difference in performance and also inspired by the recent GCD approaches, we propose to use the neighborhood information of each sample for all classes, which will give multiple positives for all data on top of their augmentations. Here, we propose not only using the nearest neighbors but also the expanded neighbors as positive pairs. Exploring the neighborhood of the samples helps to significantly increase the semantic knowledge of the latent space learned by the model.

Given an image $\mathbf{x}_i$, we obtain the l_2 normalized feature $\mathbf{z}_i = \phi(f(\mathbf{x}_i))$ by using a feature extraction backbone f and an MLP projection head ϕ. To store a subset of these normalized features, we employ a feature bank denoted as $\mathbf{F}$. To compute the neighbors of an image $\mathbf{x}_i$, we compute the cosine similarities between its normalized features and all the samples present in the feature bank $\mathbf{F}$. Finally, we select the top n samples with the highest cosine similarity values as the neighbors of $\mathbf{x}_i$. We denote these neighbors as $\{\ \mathbf{x}_j; \forall j \in N(\mathbf{x}_i)\ \}$, which are used along with the augmentations $\mathbf{x}'_i$ as positive instances, thus allowing for a comprehensive characterization of its neighborhood. Now, the neighborhood contrastive loss can be formally written as:

$$\mathcal{L}_i^N = -\frac{1}{|N(\mathbf{x}_i)|}\sum\nolimits_{q \in N(x_i)} \log \frac{\exp(\mathbf{z_i} \cdot \mathbf{z_q}/\tau)}{\sum_n 1_{[n \neq i]} \exp(\mathbf{z_i} \cdot \mathbf{z_n}/\tau)} \tag{1}$$

where $q \in N(\mathbf{x}_i)$ and $|N(\mathbf{x}_i)| = n$, number of neighbors that are taken as positives.

To enhance information aggregation, one simple strategy is to consider a greater number of nearest neighbors. However, as the neighborhood expands, it is more likely to encompass data points belonging to different classes, which hampers the desired objective of preserving class consistency. An effective alternative to capture additional target features is by leveraging the concept of expanded neighbors [34]. Here, we consider M-nearest neighbors of each neighbor in the original set, $\{\ \mathbf{x}_j$; j $\in N(\mathbf{x}_i)\ \}$. By doing so, we obtain a larger pool of expanded neighbors that offer a more comprehensive representation of the data. The index set of the expanded neighbors of image $\mathbf{x}_i$ are defined as $E_M(\mathbf{x}_i) = N(\mathbf{x}_j); \forall j \in N(\mathbf{x}_i)$. In essence, instead of directly increasing the size of the initial neighborhood, we focus on exploring the surrounding data points in a controlled manner. This way, the expanded neighbors include relevant samples that can contribute to a more robust and accurate analysis while still maintaining the principle of class consistency. The expanded neighborhood loss is formally written as :

$$\begin{aligned}\mathcal{L}_i^{EN} = -\frac{1}{|E_M(\mathbf{x}_i)|}\sum\nolimits_{q \in E_M(\mathbf{x}_i)} \lambda_{en} \, * \\ log \frac{\exp(\mathbf{z_i} \cdot \mathbf{z_q}/\tau)}{\sum_n 1_{[n \neq i]} \exp(\mathbf{z_i} \cdot \mathbf{z_n}/\tau)}\end{aligned} \tag{2}$$

where $|E_M(\mathbf{x}_i)| = M$, the number of expanded neighbors that are taken as positives, λ_{en} denotes the affinity value used for weighing the expanded neighborhood loss. While the affinity value used for all expanded neighbors are identical, it is important to note that they may not have equal significance. Upon closer inspection of the expanded neighbors, denoted as $\{\ \mathbf{x}_k$; k $\in E_M(\mathbf{x}_i)\ \}$, it becomes evident that certain neighbors might appear multiple times. For instance, a neighbor $\mathbf{x}_m$ could be the nearest neighbor for both $\mathbf{x}_h$ and $\mathbf{x}_j$, where h and j belong to the set $N(\mathbf{x}_i)$. Furthermore, the nearest neighbors themselves can also be considered as expanded neighbors. These duplicated neighbors have the potential to be semantically closer to the anchor image $\mathbf{x}_i$ by contributing to larger affinity values for these expanded neighbors, which may capture the inherent cluster structure and potentially establish stronger semantic connections. Hence, our total neighborhood loss is given by:

$$\hat{\mathcal{L}}_i^{\text{END}} = \mathcal{L}_i^N + \mathcal{L}_i^{EN} \tag{3}$$

When considering the neighbor set $N(\mathbf{x}_i)$ and expanded neighbor set $E_M(\mathbf{x}_i)$ as positives, we carefully select negatives from the feature bank $\mathbf{F}$, to avoid inadvertently classifying potential positives as negatives. To mitigate this risk, we exclusively designate samples from mini-batch B as negatives. However, we exclude the augmentation $(\mathbf{x}'_i)$ and the top n neighbors within the mini-batch from the negative set. This reduces the likelihood of mistakenly labeling potential positives as negatives within the mini-batch. This approach helps to maintain more reliable discrimination between positive and negative pairs.

For labeled data, we use supervised contrastive loss as originally given in [30], which can be written as:

$$\mathcal{L}_i^s = -\frac{1}{|M(i)|}\sum\nolimits_{q \in M(i)} \log \frac{\exp(\mathbf{z_i} \cdot \mathbf{z_q}/\tau)}{\sum_n 1_{[n \neq i]} \exp(\mathbf{z_i} \cdot \mathbf{z_n}/\tau)} \tag{4}$$

where $M(i)$ denotes a set of indices of all the other labeled samples belonging to the same class as that of $\mathbf{x}_i$ in the mini-batch B. The main difference between supervised and unsupervised contrastive loss is that the positive examples are matched by their labels in the supervised case. Hence, the final representation learning loss can be expressed as:

$$\mathcal{L}^{\text{END}} = (1-\lambda)\sum_{i \in B} \hat{\mathcal{L}}_i^{\text{END}} + \lambda \sum_{i \in B_L} \mathcal{L}_i^s \tag{5}$$

where B_L denotes the labeled subset of the mini-batch B and λ is the weight coefficient balancing the loss.

	CIFAR10 [20]	CIFAR100 [20]	Imagenet-100 [28]	Stanford Cars [19]	CUB [32]	FGVC-Aircraft [22]	Herbarium-19 [27]
$\lvert\mathcal{Y}_{\mathcal{L}}\rvert$	5	80	50	98	100	50	341
$\lvert\mathcal{Y}_{\mathcal{U}}\rvert$	10	100	100	196	200	100	683
$\lvert\mathcal{D}_{\mathcal{L}}\rvert$	12.5k	20k	31.9k	2.0k	1.5k	1.7k	5.0k
$\lvert\mathcal{D}_{\mathcal{U}}\rvert$	37.5k	30k	95.3k	6.1k	4.5k	8.9k	25.4k

Table 1. Datasets used for our experiments. Their standard split in terms of the number of labeled and unlabeled classes ($|\mathcal{Y}_{\mathcal{L}}|$, $|\mathcal{Y}_{\mathcal{U}}|$) and the number of images in the labeled and unlabeled set ($|\mathcal{D}_{\mathcal{L}}|$, $|\mathcal{D}_{\mathcal{U}}|$) are given.

3.2. Adaptive Margin Regularizer

On the classification front, inspired by [33] to train an end-to-end framework, we use a parametric classifier that uses prototype vectors that act as representatives of each class. The motivation for using an Adaptive Margin Regularizer on top of the existing classification loss is that if the prototypes of each class are well separated, the samples belonging to those classes will be better clustered.

In [12], the issue of data imbalance in image classification is tackled by repositioning the prototypes to be evenly distributed across the feature space. On the other hand, [6] proposes an adaptive approach to adjust the class prototypes, taking into consideration the data imbalance present in the training set. This technique ensures that minority class prototypes have more margin around them, reducing the likelihood of confusion between these classes with their adjacent ones. In our setting, since we don't have access to the classwise data imbalance information, we design our regularizer to adjust the prototypes adaptively by taking the current similarity between the prototypes into consideration.

A set of prototypes $\{\mathbf{c}_i\}_{i=1}^{C}$ are randomly initialized such that they correspond to each of the classes present in the dataset. C denotes the total number of classes (seen and unseen combined). The objective of the regularizer is to increase the angular separation when a pair of prototypes are very similar. We adjust the relative distance between $\mathbf{c}_i$'s such that they are at least separated by a distance greater than the mean distance between all pairs of prototypes. As the prototypes are l_2 normalized, the dot product of the prototypes denotes the cosine of the angular separation between them, minimizing which maximizes the angular separation. The adaptive margin regularizer can be written as

$$\mathcal{L}^{\text{AM}} = \frac{1}{C}\sum_{i<j}[-||\mathbf{c}_i - \mathbf{c}_j||_2^2 + d_{mean} + \Delta_{ij}] \quad \forall j \in 1,...,C \tag{6}$$

where $\Delta_{ij} = \mathbf{c}_i \cdot \mathbf{c}_j$ which tends to one when the prototypes $\mathbf{c}_i$ and $\mathbf{c}_j$ are very similar. Here, d_{mean} is the mean distance between each pair of the prototypes and it is used to stabilize the value of the regularizer. It is formally described as:

$$d_{mean}(C) = \frac{2}{C^2 - C}\sum_{i<j}[-||\mathbf{c}_i - \mathbf{c}_j||_2^2], \forall j \in 1,...,C \tag{7}$$

3.3. Final Loss

The classification objective from [33] comprises of cross-entropy loss between the soft class predictions $\mathbf{p}_i$ and ground truth labels $\mathbf{y}_i$ (for labeled data) or pseudo-labels $\mathbf{q}_i$ produced by self-distillation (for unlabeled data). The soft class predictions are obtained by applying softmax on cosine similarity between prototypes and the hidden features $\mathbf{h}_i = f(\mathbf{x}_i)$, scaled by the temperature τ_s:

$$\mathbf{p}_i^{(k)} = \frac{\exp(\mathbf{h}_i \cdot \mathbf{c}_k / \tau_s)}{\sum_{k'} \exp(\mathbf{h}_i \cdot \mathbf{c}_{k'} / \tau_s)} \tag{8}$$

where $\mathbf{h}_i$, $\mathbf{c}_k$ and $\mathbf{c}_{k'}$ are l_2 normalized. The classification loss is formally written as:

$$\mathcal{L}_{cls} = \lambda_{cls}\left(\frac{1}{|B^l|}\sum_{i\in B^l} l(\mathbf{y}_i, \mathbf{p}_i)\right) + (1-\lambda_{cls})\left(\frac{1}{|B|}\sum_{i\in B} l(\mathbf{q}'_i, \mathbf{p}_i) + \varepsilon H(\overline{\mathbf{p}})\right) \tag{9}$$

where $l(\mathbf{y}, \mathbf{p})$ refers to the cross-entropy loss between $\mathbf{y}$ and $\mathbf{p}$, and $H(\overline{\mathbf{p}})$ refers to the mean-entropy maximization regularizer [1], where $\overline{\mathbf{p}} = \frac{1}{2|B|}\sum_{i\in B}(\mathbf{p}_i + \mathbf{p}'_i)$; $\mathbf{q}'_i$ is a soft pseudo-label produced by an augmentated version of $\mathbf{x}_i$ using a model with a sharper temperature τ_t, and $\mathbf{y}_i$ denote the one hot encoding of the ground truth of $\mathbf{x}_i$; ε and λ_{cls} are the corresponding weight coefficients. The overall objective can be written as:

$$\mathcal{L}^{\text{AMEND}} = \mathcal{L}_{\text{cls}} + \lambda'\mathcal{L}^{\text{AM}} + \mathcal{L}^{\text{END}} \tag{10}$$

where λ' is the weight coefficient for the adaptive margin regularizer.

4. Experiments

Here, we perform extensive experiments to evaluate the proposed framework. First, we describe briefly the different datasets used in this work.

Data Description: We have performed experiments on three generic and four fine-grained benchmark datasets. The generic datasets used are CIFAR10, CIFAR100 [20] and ImageNet-100 [28], while the fine-grained datasets, namely Stanford Cars [19], CUB [32], FGVC Aircraft [22] and Herbarium-19 [27] are obtained from the Semantic Shift Benchmark (SSB) [31]. The presence of data imbalance is exclusively observed in the Herbarium-19 dataset, which is characterized by long-tailed data distribution. The dataset details are provided in Table 1.

Taking the training set of these datasets as $\mathcal{D}$, we sample a set of classes to belong to $\mathcal{Y}_{\mathcal{L}}$. These are referred to as the seen classes. We further sample 50% of the images for each seen class. These images then form the labeled set $\mathcal{D}_{\mathcal{L}}$, while all the other images form the unlabeled set $\mathcal{D}_{\mathcal{U}}$. In the fine-grained datasets, there is a clear semantic variation across $\mathcal{D}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{U}}$ due to the use of SSB [31]. Thus, we can evaluate if the model is learning to categorize based on clear semantic differences rather than because of underlying correlating factors present in the generic datasets. It is also assumed that there is access to a disjoint validation set $\mathcal{D}_{\mathcal{V}} = (x_i, y_i)_{i=1}^{N'} \in X \times Y_U$, where a subset of labels are masked to get the unlabeled set.

Evaluation Protocol: The model is trained using both the labeled and unlabeled data in the training set, $\mathcal{D}$, and evaluated on the unlabeled data, $\mathcal{D}_{\mathcal{U}}$. The clustering accuracy (ACC) is calculated during test time by making use of the ground truth labels y_i and the predicted cluster labels $\hat{y}_i$:

$$ACC = \max_{p \in \mathcal{P}(\mathcal{Y}_{\mathcal{U}})} \frac{1}{|\mathcal{D}_{\mathcal{U}}|} \sum_{i=1}^{|\mathcal{D}_{\mathcal{U}}|} 1\left\{y_i = p\left(\hat{y}_i\right)\right\} \tag{11}$$

where $\mathcal{P}\left(\mathcal{Y}_{\mathcal{U}}\right)$ denotes the set of all possible permutations of the class labels in $\mathcal{D}_{\mathcal{U}}$. The optimal assignment is calculated using the Hungarian algorithm [21].

The ACC is reported individually over the 'Old' classes (indicating the instances in $\mathcal{D}_{\mathcal{U}}$ that belong to the seen classes $\mathcal{Y}_{\mathcal{L}}$) and the 'New' classes (indicating the instances in $\mathcal{D}_{\mathcal{U}}$ that belong to the unseen classes in $\mathcal{Y}_{\mathcal{U}} \backslash \mathcal{Y}_{\mathcal{L}}$). Finally, the ACC over the entire $\mathcal{D}_{\mathcal{U}}$ is reported under 'All'. The model used for obtaining the performances reported follows the same protocol used by SOTA (PrompCAL [36]). The authors in [33] have employed a different protocol hence the performances are not comparable.

Implementation Details: A ViT-B-16 model pre-trained with DINO on Imagenet is used as the backbone as in the recent works [8, 30, 33], and a three-layer multi-layer perceptron (MLP) is used as the projection head, which produces a 256-dimensional feature vector as the output. We only fine-tune the last block of the ViT backbone. During training, a batch size of 128 is employed, and a feature bank with a capacity of 2048 is utilized. We follow the FIFO (First In First Out) queuing policy for storing batches of normalized features in the feature bank. The initial learning rate is set to 0.1 which is gradually annealed using a cosine scheduler with warm restarts. The weights corresponding to the representation loss and the regularizer, λ and λ' are set to 0.35 and 1.0, respectively. The affinity value λ_{en} is set to 0.1. The weight for the mean entropy maximization regularizer, ε, is set to 2. We train all the datasets for 200 epochs. For all the datasets, the number of expanded neighbors is taken as 5. The number of neighbors for all the fine-grained datasets is taken as 4. For the coarse-grained datasets, we use the number of neighbors as 5 for CIFAR10 and CIFAR100, and 6 for Imagenet-100. Though the number of neighbors for which the best performance is obtained varies slightly for different datasets, we observe (from analysis section) that the performance varies gracefully when the number of neighbors and expanded neighbors are varied. We conduct all our experiments on a single NVIDIA RTX A5000 GPU.

4.1. Comparison with Existing Approaches:

Table 2 presents the results of the proposed AMEND framework on three generic and three fine-grained datasets, along with comparisons with recent state-of-the-art methods. We observe that the proposed framework significantly outperforms the SOTA approaches on all the three fine-grained datasets, namely CUB, FGVC aircraft and Stanford Cars. For example, for the challenging Standford Cars dataset, AMEND outperforms the state-of-the-art PropmtCAL, which is also the closest competitor by 6.2%. The performance comparison on the challenging and highly imbalanced Herbarium-19 dataset is reported in Table 3. We observe that the proposed framework with its expanded neighbors and adaptive margin for handling confusing classes can effectively address the issues posed by imbalanced data, achieving the highest performance of 44.2%, which is significantly higher than the second-highest performance of 37% obtained by PromptCAL. For the generic datasets, the AMEND framework outperforms all the other approaches for the largest and most challenging ImageNet-100 dataset, while it is only second (and very close) to PromptCAL for CIFAR10 and CIFAR100 datasets.

Figure 2 shows a few qualitative results on the CIFAR10 dataset. The first (last) five columns shows examples of images which are correctly (incorrectly) classified by the proposed AMEND framework. The top two rows show

Methods	CIFAR10			CIFAR100			Imagenet-100			CUB			FGVC Aircraft			Stanford Cars		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
k-means	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3	34.3	38.9	32.1	16.0	14.4	16.8	12.8	10.6	13.8
RankStats+ [37]	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8	33.3	51.6	24.2	26.9	36.4	22.2	28.3	61.8	12.1
UNO+ [9]	68.6	98.3	53.8	69.5	80.6	47.2	70.3	95.0	57.9	35.1	49.0	28.1	40.3	56.4	32.2	35.5	70.5	18.6
ORCA	81.8	86.2	79.6	69.0	77.4	52.0	73.5	92.6	63.9	35.3	45.6	30.2	22.0	31.8	17.1	23.5	50.1	10.7
GCD [30]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3	39.0	57.6	29.9	45.0	41.1	46.9	39.0	57.6	29.9
XCON [8] (*)	96.0	97.3	95.4	74.2	81.2	60.3	82.4	90.7	78.3	52.1	54.3	51.0	47.7	44.4	49.4	40.5	58.8	31.7
DCCL [24]	96.3	96.5	96.9	75.3	76.3	70.2	80.5	90.5	76.2	63.5	60.8	64.9	-	-	-	43.1	55.7	36.2
PromptCAL [36]	**97.9**	96.6	98.5	**81.2**	84.2	75.3	83.1	92.7	78.3	62.9	64.4	62.1	52.2	52.2	52.3	50.2	70.1	40.6
AMEND (Ours)	96.8	94.6	97.8	81.0	79.9	83.3	**83.2**	92.9	78.3	**64.9**	75.6	59.6	**52.8**	61.8	48.3	**56.4**	73.3	48.2

Table 2. Performance (ACC %) of the AMEND framework and comparisons with the state-of-the-art approaches on various datasets. (*) XCON uses a batch size of 256 while other methods use a batch size of 128.

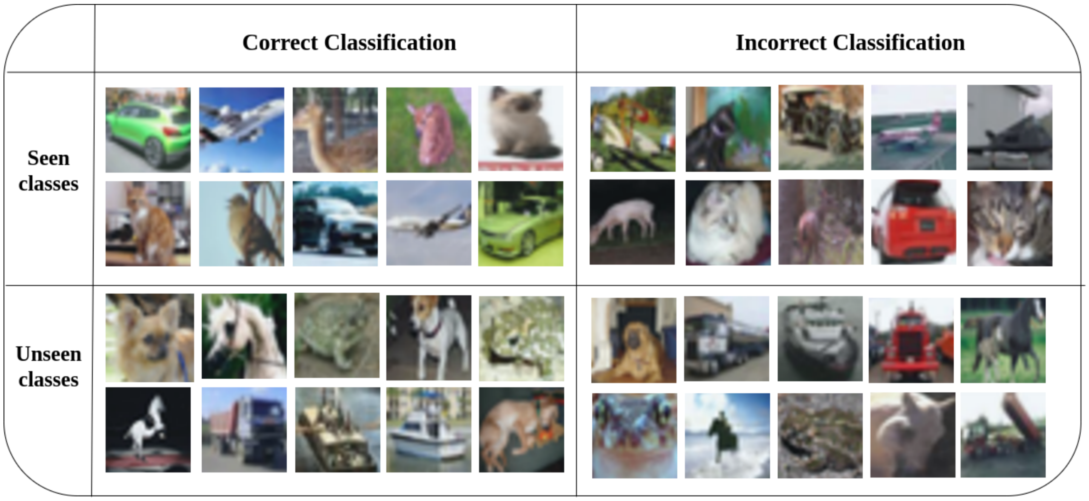

Figure 2. Few examples of correct (first five columns) and incorrect classification (last five columns) by the proposed AMEND framework for CIFAR10 dataset. The top two rows are for the seen classes and the bottom two rows are for the unseen classes.

Methods	Herbarium-19		
	All	Old	New
k-means	13.0	12.2	13.4
RankStats+ [37]	27.9	55.8	12.8
UNO+ [9]	28.3	53.7	14.7
ORCA	20.9	30.9	15.5
GCD [30]	35.4	51.0	27.0
PromptCAL [36]	37.0	52.0	28.9
AMEND (Ours)	**44.2**	**60.5**	**35.4**

Table 3. Performance of the AMEND framework on the challenging and imbalanced Herbarium-19 dataset.

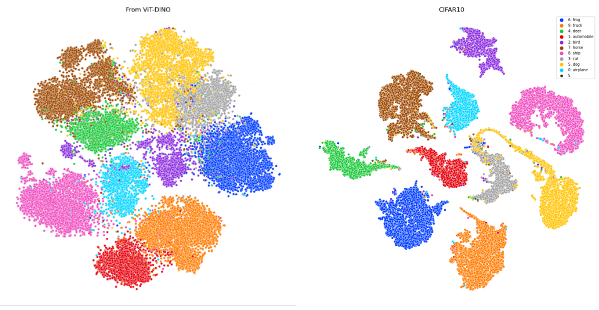

Figure 3. tSNE [29] visualization of the representations obtained from the ViT backbone for CIFAR10 for the Baseline and the proposed AMEND framework. We observe that adding the neighbors along with the adaptive margin regularizer leads to well-separated clusters, with semantically similar classes clustering together.

examples from seen classes, while the bottom two rows are examples from unseen classes. The tSNE plots for the Baseline and the AMEND framework is shown in Figure 3. We observe that the proposed framework is able to separate all the seen and unseen classes much better compared to the baseline.

Comparison with PromptCAL [36]: We observe from Tables 2 and 3 that the proposed AMEND framework outperforms all the existing approaches for five datasets, and achieves close performance to the SOTA PromptCAL for the remaining two. Here, we highlight the advantages of our approach, especially compared to PromptCAL, which lies in its simplicity and efficient implementation.

One key advantage of our method is its **memory efficiency**. We utilize a single feature bank to store 2048 features. In contrast, PromptCAL employs two separate feature banks, each with a size of 4096, for storing class embeddings and prompt embeddings, respectively. In terms of **hardware requirements**, all our experiments are conducted using a single NVIDIA RTX A5000 GPU with a capacity of 24 GB. In comparison, PromptCAL requires 48GB GPU. PromptCAL requires **warmup training** of the prompts for 200 epochs on the ImageNet-1K dataset before it can be trained on the given dataset in hand. In addition, the second training stage requires 70 epochs for generic datasets and 100 epochs for fine-grained datasets. The AMEND framework has no such requirement of warmup training and uses a single training stage for 200 epochs for all datasets. Thus, in addition to the advantages of being significantly less resource intensive, the proposed simple, yet effective AMEND framework outperforms the current state-of-the-art for five out of seven datasets, and performs at par on the remaining two.

5. Further Analysis

Here, we provide additional analysis to better understand the different components of the proposed framework. This analysis is done on CUB and Stanford Cars datasets.

Ablation Study: To evaluate the effectiveness of the different components of the proposed AMEND framework, we perform an ablation study and the results are reported in Table 4. The feature bank size is 2048, while the number of neighbors and expanded neighbors are taken as 4 and 5 respectively. We report results for the following: **Baseline** denotes the base framework which utilizes feature bank and data augmentations. We observe from the second row **(Baseline + Neighbors)** that adding the neighbors (nearest and expanded) help to improve the performance. Further incorporating the adaptive margin regularizer i.e. the complete **AMEND** framework gives the best overall performance. For the Stanford Cars dataset, inclusion of the neighbors improved the baseline performance from 52.1% to 53.2%. Finally, with the adaptive margin regularizer, the performance improves further by 3.2 %, thus achieving the state-of-the-art performance for this dataset.

Methods	**CUB**			**Stanford Cars**		
	All	Old	New	All	Old	New
Baseline	62.2	67.9	59.4	52.1	73.2	42.0
Baseline+Neighbors	63.1	72.3	58.6	53.2	74.2	43.0
AMEND (Ours)	**64.9**	**75.6**	**59.6**	**56.4**	**73.3**	**48.2**

Table 4. Ablation Study: Baseline refers to the base network which uses feature bank and augmentation. We observe from the second row that utilizing the neighbors and expanded neighbors improves the performance. Adding adaptive margin to separate confusing classes further improves the performance (third row).

Neighbors	**Expanded Neighbors**	**CUB**			**Stanford Cars**		
		All	Old	New	All	Old	New
4	5	**64.9**	**59.6**	**56.4**	56.4	73.3	48.2
5	**5**	62.8	71.2	58.6	52.6	73.1	42.7
5	**4**	62.1	71.2	57.6	54.5	72.9	45.7
5	**6**	62.7	71.7	58.2	**57.7**	**76.3**	**48.8**

Table 5. Performance of AMEND framework for different numbers of neighbors and expanded neighbors. We observe that the performance varies gradually with these hyper-parameters.

Number of Neighbors and Expanded Neighbors: Here, we analyze the effect of the number of neighbors and expanded neighbors on the classification accuracy and the results are reported in Table 5. We observe that the best overall performance for different datasets are obtained with the same number of expanded neighbors (= 5). But the number of neighbors which gave the best accuracy varied slightly with the datasets. Still, the performance is quite stable with varying the number of neighbors. Finding the optimal number of neighbors can be a future research direction.

6. Conclusion

In this work, we propose a simple, yet effective AMEND framework for the task of generalized category discovery. Specifically, we propose to incorporate neighborhood information using not only the nearest neighbors, but also expanded neighbors in the contrastive learning framework for generating robust features. Additionally, we used a class-wise adaptive margin regularizer to generate more discriminative feature representations, particularly for fine-grained datasets and confusing classes. The regularizer aimed at increasing the angular separation among the prototypes of all classes. We report results of extensive evaluation on several generic and fine-grained datasets. The AMEND framework achieves second best performance for two datasets, and establishes the new state-of-the-art performance for five datasets, including the challenging Herbarium-19 dataset. The AMEND framework offers several additional advantages in terms of less memory requirements, less hardware requirement as well as significantly lower training time as compared to the existing approaches.

7. Acknowledgement

This work is partly supported through a research grant from SERB, Department of Science and Technology, Govt. of India (SPF/2021/000118).

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 456–473. Springer, 2022. 5

[2] 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 (ICCV)*, pages 9650–9660, 2021. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, ICML'20. JMLR.org, 2020. 2

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems (NeurIPS)*, 26, 2013. 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, June 2019. 3

[6] Titir Dutta, Anurag Singh, and Soma Biswas. Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 349–364. Springer, 2020. 3, 5

[7] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9568–9577, 2021. 1, 2

[8] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *British Machine Vision Conference (BMVC)*, 2022. 1, 2, 3, 6, 7

[9] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 7

[10] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[11] 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 (ICCV)*, pages 8401–8409, 2019. 1, 2, 3

[12] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 6469–6479, 2019. 3, 5

[13] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[14] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[15] Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022. 2

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[17] 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 (NeurIPS)*, 33:18661–18673, 2020. 2

[18] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 2

[19] 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 (ICCV) workshops*, pages 554–561, 2013. 5, 6

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 6

[21] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5, 6

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[24] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7588, 2023. 2, 7

[25] S. Sarfraz, V. Sharma, and R. Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8926–8935, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. 2

[26] Yiyou Sun and Yixuan Li. Opencon: Open-world contrastive learning. In *Transactions on Machine Learning Research*, 2022. 1, 2, 3

[27] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019. 5, 6

[28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 5, 6

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

[30] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 6, 7

[31] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022. 2, 6

[32] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 5, 6

[33] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. A simple parametric classification baseline for generalized category discovery. *arXiv preprint arXiv:2211.11727*, 2022. 1, 2, 3, 5, 6

[34] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34:29393–29405, 2021. 1, 2, 4

[35] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8978–8987, October 2021. 2

[36] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3479–3488, 2023. 1, 2, 3, 6, 7

[37] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:22982–22994, 2021. 2, 7

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

[39] 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 (CVPR)*, pages 10867–10875, 2021. 1, 2

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