Proxy Anchor-based Unsupervised Learning for Continuous Generalized Category Discovery

Hyungmin Kim[1,2] Sungho Suh[3,4] Daehwan Kim[2] Daun Jeong[2] Hansang Cho[2]
Junmo Kim[1]
[1]Korea Advanced Institute of Science and Technology, Daejeon, South Korea
[2]Samsung Electro-Mechanics, Suwon, South Korea
[3]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[4]Department of Computer Science, RPTU Kaiserslautern-Landau, Kaiserslautern, Germany
{hyungmin83, junmo.kim}@kaist.ac.kr, sungho.suh@dfki.de
{daehwan85.kim, du33.jeong, hansang.cho}@samsung.com

Abstract

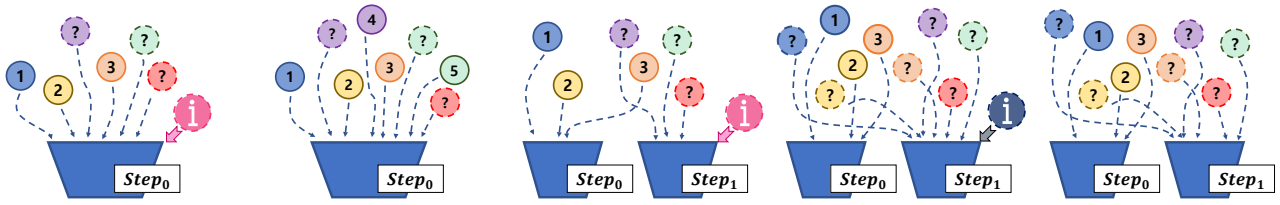
Recent advances in deep learning have significantly improved the performance of various computer vision applications. However, discovering novel categories in an incremental learning scenario remains a challenging problem due to the lack of prior knowledge about the number and nature of new categories. Existing methods for novel category discovery are limited by their reliance on labeled datasets and prior knowledge about the number of novel categories and the proportion of novel samples in the batch. To address the limitations and more accurately reflect real-world scenarios, in this paper, we propose a novel unsupervised class incremental learning approach for discovering novel categories on unlabeled sets without prior knowledge. The proposed method fine-tunes the feature extractor and proxy anchors on labeled sets, then splits samples into old and novel categories and clusters on the unlabeled dataset. Furthermore, the proxy anchors-based exemplar generates representative category vectors to mitigate catastrophic forgetting. Experimental results demonstrate that our proposed approach outperforms the state-of-the-art methods on fine-grained datasets under real-world scenarios.

1. Introduction

Deep neural networks have achieved remarkable performance in various computer vision tasks. However, current systems are still subject to constraints that are manually supervised and do not consider continual incremental categories. For extending to real-world environments, there are still gaps to catch up by overcoming the constraints and improving their abilities in fundamental tasks. Specifically, humans still perform better than machines in object cognitive and grouping skills (*e.g.* recognizing new products or clothing on shopping and categorizing undefined moving objects while driving).

Various methods have been proposed to address the limitations of the tasks by considering real-world circumstances, as presented in Figure 1 and Table 1. In detail, Novel Category Discovery (NCD) [10, 11, 43] and Generalized Category Discovery (GCD) [36, 6] aim to recognize not only pre-trained categories but also discover novel categories using a given dataset. NCD considers a disjoint dataset where labeled and unlabeled novel samples have no intersection with each other. In contrast, GCD exploits the joint set with intersected categories, making GCD a more complicated task than NCD. However, these approaches do not consider the class incremental scheme. Class incremental NCD (class-iNCD) [33, 15] has been proposed to adopt the incremental categories on NCD task, but they still focus on improving novel category discovery performance using the disjoint set, which is an unrealistic constraint. To address this issue, Grow and Merge (GM) [42] proposes a scenario that exploits the joint unlabeled dataset in the incremental novel category discovery task, which is called Continuous Category Discovery Mixed Incremental (CCD-MI). However, most existing methods require prior knowledge, such as the number of unlabeled classes for NCD and class-iNCD, or the proportion of the novel samples in the batch for CCD-MI. Such prior knowledge requirements are not enough to mimic the real-world, as we lack information about the unlabeled sets.

To overcome these constraints, we propose a novel scenario that better represents real-world circumstances by removing constraints on the available data. We assume that the given datasets are unlabeled joint sets without providing prior knowledge about the data. Employing the scenario, we propose a novel unsupervised class incremental



(a) NCD (b) GCD (c) class-iNCD (d) CCD-MI (e) CGCD

Figure 1. Comparison of existing and proposed scenarios for novel category discovery. The solid and dash-lined circles indicate labeled and unlabeled samples, respectively, with sample color depicting the class label. The circle with i denotes the number of novel classes (pink) and the proportion of novel class samples in a batch (dark blue).

Task	Method	Constraints			
		Continual learning	Assumption: $\mathcal{Y}_l \cap \mathcal{Y}_u = \emptyset$	Number of novel classes	Proportion of new class data
NCD	RankStat [10, 11], DRNCD [43]	Not considered	Required	Required	Not required
GCD	GCD [36], XCon [6]	Not considered	Not required	Not required	Not required
class-iNCD	FRoST [33], NCDwF [15]	Considered	Required	Required	Not required
CCD-MI	GM [42]	Considered	Not required	Not required	Required
CGCD	Ours	Considered	Not required	Not required	Not required

Table 1. Comparison of existing and proposed novel category discovery settings and the constraints

learning approach to simultaneously address the problems of discovering incremental novel categories and alleviating catastrophic forgetting. In addition, we focus on recognizing fine-grained objects, which is a more realistic use case for various applications in real-world applications.

The proposed method exploits a deep metric learning scheme, proxy anchor (PA) [18] that presents fast and reliable convergence and robustness against noise samples, and also considers the relations between samples. Then, we divide the unlabeled data into old and novel categories using PAs, which inherit discriminative features of old categories. The cosine similarity is measured between the PAs and the samples, and then initially separated datasets are acquired on a criterion. For further splitting, we adopt a noisy label learning scheme, and then assign the predictions of the previous model and the clustering results by a non-parametric approach to old and novel categorized samples, respectively. To mitigate the forgetting, we use a PA-based exemplar, which inherits more representative features. In the experimental results, we demonstrate that the proposed method outperforms the existing state-of-the-art in discovering novel categories and forgetting alleviation on various fine-grained datasets. Specifically, the proposed method does not require any prior knowledge and considers continual learning on unlabeled joint datasets, making it a more realistic and practical solution for real-world scenarios.

The main contributions of the proposed method can be summarized as follows.

- We introduce a novel scenario, called Continuous Generalized novel Category Discovery (CGCD), which is well-suited to tackle the challenges of discovering novel categories in real-world scenarios by removing the constraint that unlabeled data belong to only novel categories.
- We propose a novel unsupervised learning approach for incremental novel category discovery that does not require prior knowledge of the number of novel categories or the proportion of new class data.
- We present a noisy label learning approach and deep metric learning to split unlabeled data into old and novel categories, and also show mitigation of catastrophic forgetting using a deep metric-based exemplar.
- The proposed method outperforms the state-of-the-art methods in novel category discovery and forgetting mitigation on various fine-grained datasets.

2. Problem Definition

2.1. Continuous Generalized Category Discovery

As presented in Figure 1 and Table 1, various environmental schemes have been proposed to mimic real-world circumstances. Notable approaches include NCD [10, 11, 43], GCD [36], class-iNCD [33, 15], and CCD [42]. NCD considers disjoint datasets between labeled and unlabeled sets (*i.e.* $\mathcal{Y}_l \cap \mathcal{Y}_u = \emptyset$) and requires prior knowledge of the number of unlabeled categories $|\mathcal{Y}_u|$. In contrast, GCD exploits the joint set (*i.e.* $\mathcal{Y}_l \cap \mathcal{Y}_u \neq \emptyset$). Although GCD is a more challenging task than NCD, it still does not consider continuous incremental category discovery.

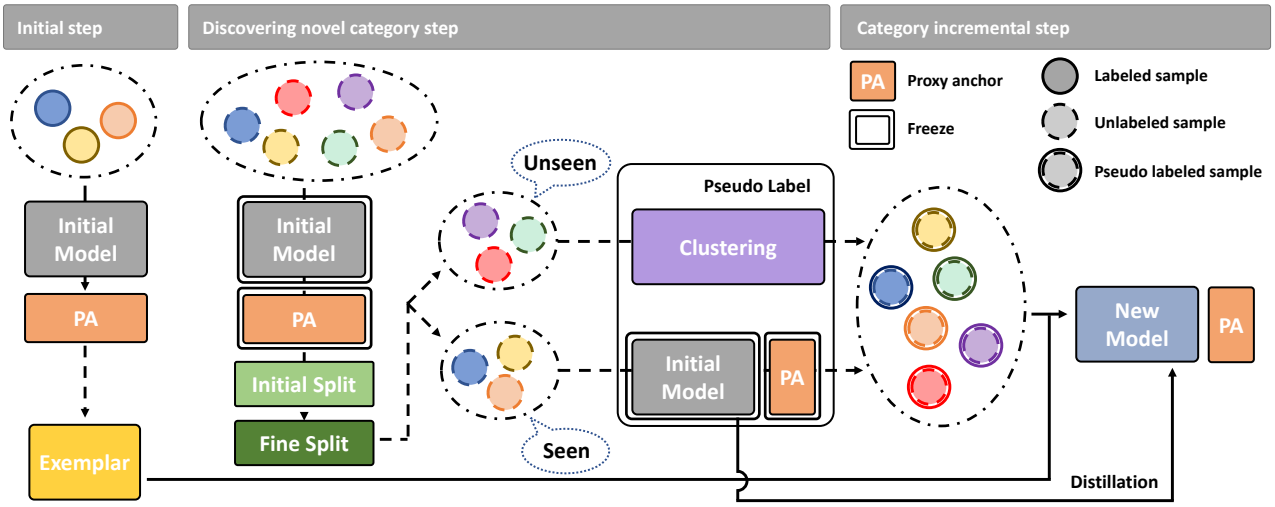


Figure 2. Overview of the proposed CGCD framework. The framework comprises three steps. The first step is that the network model is fine-tuned on the target dataset using the labeled dataset. In the second step, the discovery of novel categories is performed on the joint and unlabeled dataset, which is split into unseen and seen sets using the initial and fine split methods. Pseudo-labels are assigned to the unlabeled dataset using the previous model's predictions (seen case) and non-parametric clustering results (unseen case). In the last step, a new model is trained on the fine split dataset, which incorporates new proxy anchors based on clustering results.

Class-iNCD is an extension of NCD to the continual learning scheme. However, the method trains on the disjoint dataset under class incremental stages and has a requirement for $|\mathcal{Y}_u|$. CCD also trains on discovering novel classes under the continual learning scheme with the joint dataset. Although CCD does not require $|\mathcal{Y}_u|$, it still needs the proportion of the novel class data in a batch as prior knowledge, which is used to filter out the data.

To address these limitations, we propose a more challenging problem, named Continuous Generalized novel Category Discovery (CGCD), that is closer to real-world circumstances. In CGCD, we exploit the unlabeled joint dataset in incremental steps without providing any prior knowledge and aim to discover novel classes. This formulation is more representative of real-world scenarios, where we are not aware of the number of unlabeled categories and the characteristics of the dataset.

2.2. Setting of Continuous Generalized Category Discovery

Fine-grained datasets consist of similar objects, such as canines [16], indoor scenes [30], vehicles [25], and birds [37], in constrained circumstances. Compared to coarse datasets such as CIFAR [19] and ImageNet [20], fine-grained datasets are closer to real-world scenarios. Therefore, we focus on training and discovering novel classes with fine-grained datasets to mimic real-world circumstances better.

CGCD employs the joint unlabeled dataset in incremental steps, and Table 2 describes the dataset partitioning used in the one-time incremental category discovery. First, the set of classes is partitioned into old classes and new classes at a certain rate, for example, $8:2$. The initial step utilizes labeled dataset $\mathcal{D}^0$ consisting of old classes only. Then the following incremental step utilizes unlabeled dataset $\mathcal{D}^1$ consisting of both old classes and new classes, which reflect more realistic and challenging real-world scenarios. The key element of the proposed method is to decide whether the unlabeled data point belongs to the old classes (seen) or new classes (unseen). The samples belonging to the old classes are assigned to labeled dataset $\mathcal{D}^0$ and unlabeled dataset $\mathcal{D}^1$, and the rest of all the new class samples are assigned to $\mathcal{D}^1$. Here the choice of $8:2$ is an arbitrary example, and it is important to note that this ratio is just for data generation purposes and is not revealed to the learning methods.

Dataset	All classes	Initial step	Category incremental step	
		Old class	Old class	New class
CUB-200	200	160 (0.8)	160 (0.2)	40 (1.0)
MIT67	67	53 (0.8)	53 (0.2)	14 (1.0)
Stanford Dogs	120	96 (0.8)	96 (0.2)	24 (1.0)
FGVC Aircraft	100	80 (0.8)	80 (0.2)	20 (1.0)

Table 2. Dataset configurations for one-time incremental category scenarios. The number of classes and the proportion of data from each class in parentheses are presented. Note that the ratios in parentheses are hidden information that is not revealed to the learning methods.

3. Method

As described in Figure 2, our proposed method consists of three steps: the initial step, the novel category discov-

ery step, and the category incremental step. In the initial step, we fine-tune a pre-trained model on the labeled dataset $\mathcal{D}^0 = \{(x, y) \in \mathcal{X}_l \times \mathcal{Y}_l\}$, and obtain the embedding vector z using the model $f(\cdot)$, denoted as $z = f^0(x)$. We then use these vectors to train PAs [18] of each category, represented as $p = g^0(z)$, and also construct well-representative exemplars. In the following novel category discovery step, the given unlabeled joint datasets are denoted as $\mathcal{D}^1 = \{x | x \in \mathcal{X}_u\}$. We first separate them into old and novel categories through the initial and the fine splits. Since the separated sets are unlabeled, we pseudo-label for old and novel classes using the previous model prediction and non-parametric clustering results, respectively.

In the category incremental step, the acquired set is trained to improve the performance of discovering novel categories. To avoid catastrophic forgetting, we exploit generated features by the exemplar and feature distillation between earlier and new models. The proposed model does not require any prior knowledge, such as $|\mathcal{Y}_u|$ and the ratio of novel class samples in a batch. We evaluate the performance of the proposed method using the validation dataset, which includes all categories.

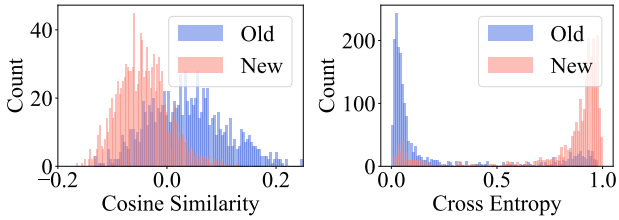
3.1. Initial Step: Fine Tune

Existing NCD methods do not account for noisy categories, such as those categorized from old to novel or from novel to old, which can impair novel discovery performance and accumulate errors in the continuous procedure. To address these limitations, in this work, we propose a novel approach that leverages the benefits of PA to complement and improve the existing approaches. PA is a metric learning method that combines proxy- and pair-based methods to achieve rapid and reliable convergence, and robustness against noisy samples, and considers relations between data to extract meaningful semantic information.

Following the method, the embedding vector z from the initial model f^0 is trained to map to each proxy anchor $p = g^0(z)$. Let the set of all proxy anchors as P^0 in the labeled data $\mathcal{D}^0$. In this manner, the number of proxy anchors of $\mathcal{D}^0$ is the number of classes of the labeled set (*i.e.* $|P^0| = |\mathcal{Y}_l|$) in the initial step. We train the model and proxy anchors using the following loss function defined in [18]:

$$\mathcal{L}^0_{pa}(Z^0) = \frac{1}{|P^{0+}|} \sum_{p \in P^{0+}} \log\left(1 + \sum_{z \in Z_p^{0+}} e^{-\alpha(s(z,p)-\delta)}\right) + \frac{1}{|P^0|} \sum_{p \in P^0} \log\left(1 + \sum_{z \in Z_p^{0-}} e^{\alpha(s(z,p)+\delta)}\right) \tag{1}$$

where $\delta > 0$ is a margin and $\alpha > 0$ is a scaling factor. The function $s(\cdot,\cdot)$ indicates the cosine similarity score. P^{0+} represents same class PAs(*e.g.*positive) in the batch. Each proxy p divides the set of embedding vector Z^0 as Z_p^{0+} and $Z_p^{0-} = Z^0 - Z_p^{0+}$. Z_p^{0+} denotes the same class embedding points with the proxy anchor p. The first term aims to pull p and its dissimilar but hard positive data together, while the last term is to push p and its similar but hard negatives apart.



(a) Initial Split (b) Fine Split

Figure 3. Separation results on joint unlabeled datasets

3.2. Discovering Novel Categories Step

Separation: In this procedure, we aim to split the given joint dataset $\mathcal{D}^1$ into the novel and old categories without any prior knowledge. We conduct this task in two stages: initial split and fine split. In the initial split, we compute the cosine similarity between p and each embedding vector $z_i \in Z^1$, where $z_i = f^0(x_i)$ and $x_i \in D^1$. Because the set of proxy anchors P^0 represents the old categories, we classify a sample to the old class if the maximum similarity score of z_i is larger than a threshold ϵ. We set $\epsilon = 0$ since it is the median of the score ranges. The initial split is defined as:

$$\tilde{y}_i = \begin{cases} 0, & \text{if } \max\limits_{p \in P^0}(s(z_i, p)) \geq \epsilon = 0 \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

To acquire a cleaner novel and old dataset, we propose a noisy labeling scheme, fine split, which involves an iterative training of a simple multilayer perceptron (MLP) based classifier $m(\cdot)$ on the binarized dataset. The initial split results in noisy and inaccurate separation, as shown in Figure 3 (a). Among them, only the data on both ends of the spectrum are assumed the clean and utilized to train the classifier. The loss of split network $m(\cdot)$ is defined as follows:

$$\mathcal{L}_{sp} = -\mathbb{E}_{z_c \in Z_c^1}[\tilde{y}_c \log(m(z_c)) + (1 - \tilde{y}_c) \log(1 - m(z_c))] \tag{3}$$

where z_c denotes clean embedding vectors, and Z_c^1 represents the set of clean vectors. $\tilde{y}_c$ indicates the pseudo label of the clean data. After the warm-up training, the classifier is trained with re-assigned pseudo labels and the more cleaned data, which is divided with the Gaussian mixture model (GMM). Consequently, the Figure 3 (b) shows the cleaner separated results.

Pseudo-labeling: After the separation, both the old $\mathcal{D}^1_{old}$ and the novel categories $\mathcal{D}^1_{new}$ are still unlabeled. Thus, we use pseudo-labels to assign labels to each sample. For $\mathcal{D}^1_{old}$, we use the predictions of the previous model and proxy anchors to assign pseudo-labels. In contrast, we use a

non-parametric clustering approach named Affinity propagation [8] to assign pseudo-labels to $\mathcal{D}^1_{new}$. In this manner, our proposed approach does not require any prior knowledge. Finally, from the clustering results, we obtain an estimate of the number of novel categories, denoted as $|\hat{\mathcal{Y}}_n|$.

3.3. Category Incremental Step

Training modified model and PAs: To improve the performance of discovering novel categories, we modify the model since the previous model has PAs only for $|\mathcal{Y}_l|$ classes and cannot categorize novel classes. We add new p for $|\hat{\mathcal{Y}}_n|$ classes, increasing the total number of PAs like $|P^1| = |\mathcal{Y}_l| + |\hat{\mathcal{Y}}_n|$. The loss function to train the modified model f^1 and PA $p = g^1(z)$ is reformulated as follows:

$$\begin{aligned}\mathcal{L}^1_{pa}(Z^1) = &\frac{1}{|P^{1+}|}\sum_{p\in P^{1+}}\log\left(1+\sum_{z\in Z^{1+}_p} e^{-\alpha(s(z,p)-\delta)}\right)\\ &+\frac{1}{|P^1|}\sum_{p\in P^1}\log\left(1+\sum_{z\in Z^{1-}_p} e^{\alpha(s(z,p)+\delta)}\right)\end{aligned} \tag{4}$$

Avoiding forgetting: In the continual learning scheme, it is essential to alleviate catastrophic forgetting. We adopt feature replay, which leverages the PA information belonging to old categories. Each well-trained p inherits the representation power for each category. We employ each p to generate features by following the Gaussian distribution $\mathcal{N}(p^0, \sigma^2), p^0 \in P^0$. The number of generated features is determined based on data balancing, for example, the number of newly categorized samples in a batch. The generated features are concatenated into a batch, and the model and PAs are trained using the following loss function:

$$\mathcal{L}^1_{ex}(\tilde{Z}) = \mathcal{L}^1_{pa}(\tilde{Z}), \qquad \tilde{Z} = \{\tilde{z} \sim \mathcal{N}(p^0, \sigma^2)\} \tag{5}$$

Also, we utilize the distillation of the extracted embedding vectors from the present f^1 and the previous model f^0. The distillation loss $\mathcal{L}_{kd}$ is described as follows:

$$\begin{aligned}\mathcal{L}^1_{kd}(z_o) &= -\mathbb{E}_{z_o\in Z^1_{old}}\|z^0_o - z_o\|_2\\ &= -\mathbb{E}_{x_o\in D^1_{old}}\|f^0(x_o) - f^1(x_o)\|_2\end{aligned} \tag{6}$$

where z^0 represents the embedding vector for fixed previous feature network f^0. Z^1_{old} denotes seen data from $Z^1 = \{Z^1_{old} \cup Z^1_{new}\}$.

In conclusion, the loss consists of three different losses in the continuous category discovery step. One loss is for training the PAs and the model on $\mathcal{D}^1$, the others are to avoid forgetting by using generated features and knowledge distillation. $\mathcal{L}^1$ is described as follows:

$$\mathcal{L}^1 = \mathcal{L}^1_{pa}(Z^1) + \mathcal{L}^1_{ex}(\tilde{Z}) + \mathcal{L}^1_{kd}(z_o) \tag{7}$$

4. Experimental Results

4.1. Implementation Details

We utilized the widely used augmentation techniques, including random crop after padding and random horizontal flip. All the experiments were trained for 60 epochs using AdamW optimizer with weight decay set to 0.0001. The initial learning rate was set to 0.0001 for the model $f(\cdot)$, while for the PAs, it was set to 0.01. The learning rate was decayed by a factor of 0.5 every five epochs. We used the threshold ϵ only once for the initial split to divide the set into old and novel categories, and we set it to 0 for all the datasets and networks. For fine split, we used an MLP-based network architecture that consists of two dense layers with a batch normalization layer. The model was trained for three epochs using AdamW with a learning rate of 0.0001. The hyperparameters for PAs, α and δ, were set to 32 and 0.1, respectively.

For fair comparisons of various methods, such as NCD, class-iNCD, and CCD, we follow the hyperparameters and the network architectures of the original implementations, referring to the papers for details. All the reported performances are average results over three runs.

4.2. Evaluation Metrics

We evaluate the methods using metrics based on the cluster accuracy measurement, called Hungarian assignment algorithm [21]. The evaluation metric is defined as follows:

$$\mathcal{M}^t = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\mathbb{I}(y_i = h^*(y^*_i)) \tag{8}$$

where $|\mathcal{D}|$ is the size of the validation dataset $\mathcal{D}$ and h^* is the optimal assignment. So, $\mathcal{M}^t$ measures the cluster accuracy at step t on the $\mathcal{D}$. In this manner, $\mathcal{M}_{all}$ and $\mathcal{M}_o$ indicate the cluster accuracy metrics of the whole and old categories using $\mathcal{M}^t$, respectively. Furthermore, we employ two more metrics that, $\mathcal{M}_f$ and $\mathcal{M}_d$, which are proposed on GM [42] and described as follows:

$$\mathcal{M}_f = \max_t\{\mathcal{M}^0_o - \mathcal{M}^t_o\}, \tag{9}$$

$$\mathcal{M}_d = \frac{1}{|T|}\sum_{i=T}\mathcal{M}^i_n. \tag{10}$$

where $\mathcal{M}^0_o$ and $\mathcal{M}^t_o$ are the old class cluster accuracy at the initial step and category incremental t step, respectively. So $\mathcal{M}_f$ measures the maximum forgetting values for old categories in the entire step and should be sufficiently low; if not, the method is not valuable in practical applications. In Equation (10), $|T|$ is the number of increased steps, and $\mathcal{M}_d$ evaluates the averaged performance of novel category discovery in unlabeled joint datasets in each step. It means

Method	CUB-200				MIT67				Stanford Dogs				FGVC aircraft			
	$\mathcal{M}_{all}\uparrow$	$\mathcal{M}_o\uparrow$	$\mathcal{M}_f\downarrow$	$\mathcal{M}_d\uparrow$	$\mathcal{M}_{all}\uparrow$	$\mathcal{M}_o\uparrow$	$\mathcal{M}_f\downarrow$	$\mathcal{M}_d\uparrow$	$\mathcal{M}_{all}\uparrow$	$\mathcal{M}_o\uparrow$	$\mathcal{M}_f\downarrow$	$\mathcal{M}_d\uparrow$	$\mathcal{M}_{all}\uparrow$	$\mathcal{M}_o\uparrow$	$\mathcal{M}_f\downarrow$	$\mathcal{M}_d\uparrow$
Supervised	61.69	45.83	23.32	16.63	55.56	40.90	19.52	18.34	64.26	47.57	25.43	17.01	64.62	48.17	26.24	18.12
DRNCD [43]	9.80	10.47	58.51	34.24	26.99	27.67	49.07	38.91	16.39	10.34	65.83	63.36	18.73	19.50	57.05	45.63
FRoST [33]	18.19	17.34	12.18	17.20	23.21	23.55	15.96	24.74	22.62	22.23	14.29	26.06	32.61	33.53	15.19	27.69
GM [42]	6.43	6.57	39.82	5.92	16.52	16.77	42.26	15.50	5.99	5.98	50.36	5.98	12.00	11.63	53.35	13.46
Ours	**54.75**	**58.80**	**15.47**	**40.90**	**54.45**	**64.23**	**10.64**	**18.58**	**66.25**	**76.15**	**6.77**	**30.04**	**37.28**	**41.60**	**14.66**	**20.04**

Table 3. Comparison results under continuous generalized categorized discovery scenario. The results present the mean over three runs.

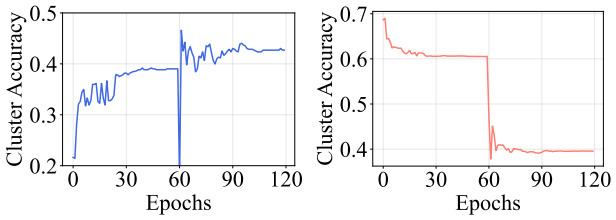
Figure 4. Qualitative evaluation results of the proposed method using the CUB-200 dataset on ResNet-18. The first five columns with blue boxes denote well-clustered examples. The last two columns represent failed prediction results, including example images with purple boxes denoting hard negatives and those with red boxes indicating incorrect categorization.

the higher method, the more appropriate method in real-world applications.

4.3. Comparison with State-of-the-arts Methods

We conducted a series of experiments to compare the cluster accuracy performance of our proposed method with state-of-the-art approaches, as presented in Table 1. In the experiments, we excluded the GCD task, including XCon, as it focuses on discovering novel categories blended with labeled and unlabeled datasets and evaluates unknown discovery performance on only train datasets, not validation datasets, in their papers. Therefore, we compared our proposed method with other approaches, including Dual rank NCD (DRNCD) [43], FRoST [33], and GM, which are the representative methods that record the state-of-the-art results of each task.

We first evaluated the one-step incremental category setting and reported the comparison results in Table 3. The supervised method is the setting of a supervised continual learning manner, literally. We observed catastrophic forgetting in supervised learning. DRNCD is one of the NCD approaches and recorded the outstanding performances of $\mathcal{M}_d$ on MIT67, Stanford Dogs, and FGVC aircraft. However, the method requires prior knowledge of the number of novel categories and does not identify specific classes, but only knows whether the samples are included in the novel category or not. Thus, the results are not regarded as outperforming. Instead, the method shows the highest results of $\mathcal{M}_f$, which means that the method focuses on learning novel categories without considering the prevention of forgetting previous categorical knowledge. In this regard, NCD is not sufficient to extend novel class incremental learning schemes. FRoST showed competitive results of discovering novel classes and decreased forgetting results $\mathcal{M}_f$ on all datasets compared to DRNCD. Nevertheless, considering the required knowledge, the other metrics results, $\mathcal{M}_{all}$ and $\mathcal{M}_o$ were not competitive. GM proposed the CCD setting, which is the most similar to CGCD, in the perspective of the unknown number of novel categories and discovering novel and old categories on new unlabeled datasets. However, the method requires a crucial parameter, which is the ratio of novel category samples on the new dataset. Without considering the ratio, GM recorded the lowest results of $\mathcal{M}_{all}$, $\mathcal{M}_o$, and $\mathcal{M}_d$.



(a) Novel category discovery M_d (b) Catastrophic forgetting M_f

Figure 5. Performances of novel category discovery and forgetting on two-step incremental novel categories experiment

The proposed method showed outstanding performance on the various datasets without requiring any prior information about new incoming unlabeled datasets. Our method recorded the second-best $\mathcal{M}_f$ on the CUB-200 dataset and the best $\mathcal{M}_f$ on the other datasets, such as MIT67, Dogs, and FGVC aircraft, by the effects of PA-based exemplar. On $\mathcal{M}_d$, the method was also competitive and compared to the GM, which has the most similar setting to our method.

To evaluate the proposed method qualitatively, we clustered the evaluation dataset using the CUB-200 dataset. In Figure 4, our method well-discovered novel categories and clustered them correctly. Each row is clustered into the same category, and the classes are novel categories on the evaluation dataset. The left five columns are well-clustered, while the last two are not. The sixth-column images are still reasonable, but the last-columns are the worst cases.

4.4. Two-step Novel Category Discovery

We present a two-step incremental category discovery experiment on the CUB-200 dataset using ResNet-18. The dataset configurations are more complicated as the datasets are joint sets in incremental steps. The initial step, the first incremental step, and the second incremental step have novel classes in each step, at a rate of $8\colon 1\colon 1$, respectively. Each step has its dataset and is indicated as $\mathcal{D}^0$, $\mathcal{D}^1$, and $\mathcal{D}^2$. The samples belonging to novel labeled classes in the initial step are assigned to $\mathcal{D}^0$, $\mathcal{D}^1$, and $\mathcal{D}^2$ at a rate of $8\colon 1\colon 1$. Similarly, the samples belonging to novel classes in the first incremental step are assigned to $\mathcal{D}^1$, and $\mathcal{D}^2$ at a rate of $8\colon 2$. Finally, the rest of the samples are assigned to $\mathcal{D}^2$.

Figure 5 describes the performance of the experiments. Since each incremental step is trained for 60 epochs, there are deep drops at the $60th$ epoch when new PAs are added. In addition, the cluster accuracy of the old categories, which belong to the $\mathcal{D}^0$, decreases by about 20%. The reason is that the number of old categories in the second incremental step is increased compared to the first step. The exemplar cannot focus on only generating the features of $\mathcal{D}^0$. However, the performance of $\mathcal{M}_d$ increased steadily.

4.5. Ablation Study

Effectiveness of separations: We conducted an ablation study to show the effectiveness of our proposed splitter, which consists of a combination of two different methods: the cosine similarity score and a binary splitter using the noise label approach. The ablation experiments were designed such that one used only the cosine similarity score, and another used both methods. As shown in Table 4, the approach using both methods presents improvements in both old and novel discovery performances. The results reveal the effectiveness of noise labeling for data separation.

Network	Fine Split	Metric			
		$\mathcal{M}_{all}\uparrow$	$\mathcal{M}_o\uparrow$	$\mathcal{M}_f\downarrow$	$\mathcal{M}_d\uparrow$
ResNet-18	without	53.32	57.63	16.56	36.47
	with	54.75	58.80	15.47	40.90
ResNet-50	without	66.53	71.07	9.12	48.76
	with	68.09	71.75	8.44	53.79
ViT-B-16	without	70.22	73.02	10.75	59.24
	with	72.51	74.28	9.49	65.60

Table 4. Ablation study for the proposed fine split on the CUB-200 using three different network architectures, including ResNet-18 and ResNet-50 pre-trained on ImageNet, and ViT-B-16 pre-trained on DINO-ImageNet. The results present the mean over three runs.

Network	Exemplar	Metric			
		$\mathcal{M}_{all}\uparrow$	$\mathcal{M}_o\uparrow$	$\mathcal{M}_f\downarrow$	$\mathcal{M}_d\uparrow$
ResNet-18	without	38.24	37.31	36.88	41.89
	Data mean and std.	40.23	40.56	33.63	37.22
	Proxy anchors	54.75	58.80	15.47	40.90

Table 5. Effectiveness of the proposed PA-based exemplar on the CUB-200 dataset using ResNet-18 pre-trained on ImageNet

Effectiveness of Proxy anchor exemplar: To mitigate catastrophic forgetting, various approaches, such as replay [2], prototype [31, 33], and pseudo-latents [15], have been proposed. Most of these methods exploit the computed average of feature embedding vectors or input data-driven values. However, we propose a novel PA-based exemplar approach and evaluate its efficiency. As shown in Table 5, the method without the exemplar recorded the highest novel discovery performances but also showed the highest forgetting. Adopting a general exemplar approach using mean and standard deviation values from the former dataset, $\mathcal{M}_f$ slightly decreased, but $\mathcal{M}_d$ is the lowest. However, our proposed method with the PA-based exemplar recorded the best $\mathcal{M}_f$ and competitive $\mathcal{M}_d$. We analyze that the PAs have representatives of each cluster since PA inherits the relation between data to data and then representative figures of each class. Hence, our PA-based method leads to mitigating forgetting, and we confirm that it is a proper approach.

Robustness of class and sample blending ratio variants: In general, the capability to recognize novel categories largely depends on a powerful and well-trained initial model with the target datasets. The more classes and

Ratio	Metric			
$\mathcal{Y}_{old}$: $\mathcal{Y}_{new}$	$\mathcal{M}_{all}\uparrow$	$\mathcal{M}_{o}\uparrow$	$\mathcal{M}_{f}\downarrow$	$\mathcal{M}_{d}\uparrow$
9 : 1	61.34	63.22	11.83	44.73
8 : 2	54.75	58.80	15.47	40.90
7 : 3	52.66	58.42	14.67	39.50
6 : 4	48.78	57.90	15.16	35.53
5 : 5	45.28	62.41	11.30	28.55

Table 6. Qualitative evaluation by changing the ratio of classes and samples on the CUB-200 using ResNet18 pre-trained on ImageNet

samples are included in the initial training dataset $\mathcal{D}^0$, the better the model learns representative features to fit the sets. To evaluate the robustness of the proposed model, we conducted experiments with variants of the number of samples and classes in $\mathcal{D}^0$. As described in Table 6, decreasing the number of classes and data in the labeled set decreased the discovery of novel classes and clustering accuracies as the number of unlabeled novel data increased. On the other hand, catastrophic forgetting could increase since the number of novel data increases. However, forgetting was maintained within a reasonable boundary, indicating the effectiveness of our PA-based exemplar. The results suggest that our method has the robustness of the variants.

5. Related work

5.1. Novel Category Discovery

NCD techniques have been proposed to classify data with various constraints on unlabeled data. One category of the methods presented pre-training the model on the labeled set and fine-tuning it on the unlabeled set using unsupervised clustering losses [41, 13, 12, 23, 24]. Another category assumed the availability of both the labeled and unlabeled data, and trained networks jointly with a labeled novel class loss within the semi-supervised scheme [10, 44, 45, 14, 7, 43]. Han *et al*. [11] proposed transferring knowledge from labeled to unlabeled data using ranking statistics in the joint learning stage. Recently, GCD [36] and XCon [6] tackled the more realistic scenario of joint datasets and distinguished known and unknown classes using prior knowledge. However, these approaches did not consider the continual learning scheme. To address the limitation, FRoST [33] and NCDwF [15] froze feature extractors and added the second head for each novel class, as much as the given number of novel categories. However, the methods employed disjoint sets. GM [42] proposed to consider novel category discovery on the joint datasets, but still require prior knowledge, such as the proportion of novel samples.

5.2. Image Retrieval

Most of the image retrieval methods have utilized metric learning and can be categorized into two approaches. Pair-based methods exploited contrastive loss [3, 5, 9] and triplet loss [34, 38], that pull together data pairs in the same class and push apart those in different classes. Multiple data-based [35, 28] methods proposed considering the relations between multiple data. Entire data-based approaches [40, 39] presented considering all data in a batch, leveraging fine-grained semantic relations between them while requiring high computation costs and slow convergence. In contrast, proxy-based methods [26, 29, 1] employed fewer proxies than the training set, reducing training complexity. While these methods improved training convergence, they did not consider data-to-data relations, as each data was associated with its proxy. PA [18] inherited the strength of pair- and proxy-based methods, achieving fast and reliable convergence, robustness opposing noisy data, and leveraging rich data-to-data relations.

5.3. Noise Label

Recently proposed methods for learning with noisy labels have highlighted the importance of discriminating between clean and noise-labeled data to improve performance. DivideMix [22] used GMM to distinguish between clean and noisy labeled data and treated the latter as unlabeled for semi-supervised learning. AugDesc [27] employed data augmentation to enhance the differentiation between clean and noisy labeled data, while INCV [4] introduced cross-validation to separate clean data from noisy training data. SplitNet [17] leveraged a compact network to perceive the difference between clean and noisy labels, improving model performance by more accurately differentiating noise.

6. Conclusion

In this paper, we presented a novel continual learning scenario, considered NCD on the unlabeled joint datasets without any prior knowledge of the dataset. Our framework utilized PAs to split known and novel categories, resulting in well-clustered and well-pseudo-labeled categories that mitigate catastrophic forgetting. We further refined the splitting of the dataset by adopting a noise labeling scheme. Our proposed approach outperformed existing state-of-the-art methods regarding novel category discovery and forgetting. While DeepDPM [32] has recently shown outstanding performance on non-parametric clustering tasks, we believe that our proposed method can achieve even better performance by adopting better clustering manners. In future work, we plan to evaluate our method by adopting a better clustering manner.

Acknowledgements

This work was funded by Samsung Electro-Mechanics and was partially supported by Carl-Zeiss Stiftung under the Sustainable Embedded AI project (P2021-02-009).

References

[1] Nicolas Aziere and Sinisa Todorovic. Ensemble deep manifold similarity learning using hard proxies. In *CVPR*, pages 7299–7307, 2019.

[2] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *CVPR*, pages 8218–8227, June 2021.

[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a siamese time delay neural network. In *NeurIPS*, volume 6, 1993.

[4] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, pages 1062–1070, 2019.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

[6] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *BMVC*, 2022.

[7] Enrico Finia, Enver Sangineto, Stephane Lathuiliere, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021.

[8] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[9] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742, 2006.

[10] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020.

[11] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *TPAMI*, 2021.

[12] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019.

[13] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019.

[14] Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *ICCV*, 2021.

[15] K. J. Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N. Balasubramanian. Novel class discovery without forgetting. In *ECCV*, page 570–586, 2022.

[16] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop on Fine-Grained Visual Categorization*, 2011.

[17] Daehwan Kim, Kwangrok Ryoo, Hansang Cho, and Seungryong Kim. Splitnet: Learnable clean-noisy label splitting for learning with noisy labels, 2022.

[18] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *CVPR*, June 2020.

[19] Alex Krizhevsky and Geoff Hinton. Learning multiple layers of features from tiny images, 2009.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[21] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, March 1955.

[22] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020.

[23] Yu Liu and Tinne Tuytelaars. Residual tuning: Toward novel category discovery without labels. In *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[24] Yu Liu and Tinne Tuytelaars. Residual tuning: Toward novel category discovery without labels. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[25] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013.

[26] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *CVPR*, pages 360–368, 2017.

[27] Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *CVPR*, pages 8022–8031, 2021.

[28] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016.

[29] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *CVPR*, pages 6450–6458, 2019.

[30] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009.

[31] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017.

[32] Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. Deepdpm: Deep clustering with an unknown number of clusters. In *CVPR*, pages 9861–9870, June 2022.

[33] Subhankar Roy, Mingxuan Liu, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Class-incremental novel class discovery. In *ECCV*, 2022.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[35] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *NeurIPS*, volume 29, 2016.

[36] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022.

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

[38] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, pages 1386–1393, 2014.

[39] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pages 5022–5030, 2019.

[40] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *CVPR*, pages 5207–5216, 2019.

[41] Zhaoyang Lv Yen-Chang Hsu and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018.

[42] Xinwei Zhang, Jianwen Jiang, Yutong Feng, Zhi-Fan Wu, Xibin Zhao, Hai Wan, Mingqian Tang, Rong Jin, and Yue Gao. Grow and merge: A unified framework for continuous categories discovery. In *NeurIPS*, 2022.

[43] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, 2021.

[44] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luoa nd Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, 2021.

[45] 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 *CVPR*, 2021.