On-the-fly Category Discovery

Ruoyi Du[1], Dongliang Chang[1], Kongming Liang[1]*, Timothy Hospedales[2], Yi-Zhe Song[3], Zhanyu Ma[1]
[1]Beijing University of Posts and Telecommunications, China
[2]University of Edinburgh, UK [3]University of Surrey, UK
{duruoyi, changdongliang, liangkongming, mazhanyu}@bupt.edu.cn,
t.hospedales@ed.ac.uk, y.song@surrey.ac.uk

Abstract

*Although machines have surpassed humans on visual recognition problems, they are still limited to providing closed-set answers. Unlike machines, humans can cognize novel categories at the first observation. Novel category discovery (NCD) techniques, transferring knowledge from seen categories to distinguish unseen categories, aim to bridge the gap. However, current NCD methods assume a transductive learning and offline inference paradigm, which restricts them to a pre-defined query set and renders them unable to deliver instant feedback. In this paper, we study on-the-fly category discovery (OCD) aimed at making the model instantaneously aware of novel category samples (i.e., enabling inductive learning and streaming inference). We first design a hash coding-based expandable recognition model as a practical baseline. Afterwards, noticing the sensitivity of hash codes to intra-category variance, we further propose a novel **S**ign-**M**agnitude d**I**sentang**LE**ment (SMILE) architecture to alleviate the disturbance it brings. Our experimental results demonstrate the superiority of SMILE against our baseline model and prior art. Our code is available at https://github.com/PRIS-CV/On-the-fly-Category-Discovery.*

1. Introduction

Deep models are well known for beating humans in visual recognition [13]. However, this is just a victory of specialist models over generalist humans – existing vision recognition models are mostly closed-set experts. Given a defined category set, huge datasets are gathered and annotated, and then, deep models trained with the annotated data can easily handle such an in-category recognition due to their great fitting ability. However, these models are arguably only *learning to memorize* in that they are restricted to the defined category set and are incapable of modeling novel categories. Although paradigms like open set recognition [9] aim to filter out the out-of-category samples, simply rejecting them is not satisfactory. For humans, visual recognition is far beyond a closed-set problem – instead of *learning to memorize*, we *learn to cognize*. In particular, given samples containing novel categories, we can not only tell which are novel but we can also tell which may share the same novel category. *E.g.,* even you have never seen "hedgehogs", you can easily realize that they differ from other creatures you have seen before and realise that multiple hedgehog images belong to the same category, even if you don't know the name.

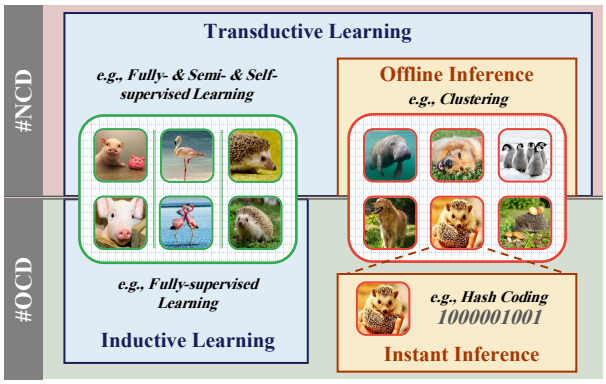


Figure 1. Comparison of the conventional NCD setting and the proposed OCD setting. (a) NCD adopts transductive learning and offline inference. (b) OCD removes the pre-defined query set assumption and conducts inductive learning and instant inference.

To bridge the gap, a rising field named novel category discovery (NCD) [11] attaching increasing attention. With a labelled support set of seen categories and an unlabeled query set containing unseen ones, NCD aims at cognizing unseen categories by splitting the query set into several groups with the same latent category. As shown in Figure 1, existing NCD works [7, 10, 15, 37] mostly fall into a transductive learning and offline inference procedure. Specifically, a visual feature encoder is first trained with the support set via supervised learning and the query set via unsupervised or semi-supervised learning. After that, clustering

*Corresponding Author

techniques are applied to the encoded visual features to obtain category clusters.

Although convincing performance has been obtained, two restrictive assumptions still hinder the real-world application of NCD approaches under the current setting. (i) Firstly, the query set is visible and required during training, which makes the model specialized to the pre-defined query set and less capable of dealing with truly novel samples. (ii) Secondly, the query set is batch processed offline during inference. Therefore these models are not practical in online scenarios where new data occurs in a stream and instant feedback on each instance is required.

To approach a more realistic scenario, we put forward the problem of on-the-fly category discovery (OCD) that removes the assumption of a pre-defined query set (Figure 1). In particular, we keep the seen/unseen split of datasets, and make samples of the unseen query set unavailable during training and only individually visible during test. The goal of OCD is learning to recognise seen categories and to cognize unseen categories – both in an inductive manner that can be applied online. We follow the setting of generalized category discovery (GCN) [30] where both seen and unseen categories appear in the query set.

Next, we introduce a new recognition paradigm for OCD along with a baseline model. Instead of adopting cross-entropy loss during training for fully supervised learning, we choose supervised contrastive learning [16] that works in embedding space. Thus, we directly optimize and obtain discriminative visual features rather than probability outputs within a fixed prediction space. To meet the need for instant feedback, cluster-based techniques are no longer practical during inference. To this end, we take the binarized feature embeddings as hash-like category descriptors, and samples with the same descriptor can be regarded as sharing the same latent category. In this way, the model can individually recognize each novel sample, like us humans.

Afterward, we observe a challenge of OCD – the hash-like descriptor is extremely sensitive to intra-category variance, especially for fine-grained categories. *E.g.*, for the CUB-2011-200 dataset [31], ~1500 different 12-dimension hash codes are generated for ~4000 birds from only 200 categories. To address this, we contribute a novel **S**ign-**M**agnitude d**I**sentang**LE**ment (SMILE) architecture to alleviate the negative influence of intra-category variance. Specifically, we infer the signs and the magnitudes of feature embeddings with two separate branches and only the sign branch is used during inference. The intuition behind this is that, since deep neural features respond to abstract semantics (e.g., colors, textures, shapes), the sign branch should encode whether a semantic feature corresponds to this category, and the magnitude branch indicates the expression level of the semantic feature on the current sample. In summary, the magnitude branch should model intra-category variance, and the sign branch inter-category variance. Experiments on three widely used classification datasets and three fine-grained classification datasets demonstrate the superiority of SMILE over our baseline and prior art.

2. Related Work

Open-set Recognition A relevant pioneer of visual recognition in real-world scenarios is open-set recognition (OSR) [9] with a history of nearly a decade [26]. OSR supposes novel categories appearing in the testing set and aims at rejecting them during inference. At the very beginning, Scheirer *et al.* [26] provided a preliminary solution for OSR by introducing open space risk and proposing an "1-vs-set machine" to define the open-set margin. Afterwards, relevant research has mainly followed two trends of generative and discriminative models. For generative model-based OSR [8, 17, 23], a generative model is often employed to synthesize samples from unseen categories, and the open-set decision boundary can be learned in a supervised manner. As for discriminative model-based approaches, they adjust the decision space by directly modeling the open-set margin via SVM [3,27], sparse representation [35], distance measurement [1], *etc*. However, current OSR models are still limited to awareness of unseen categories, while we NCD and OCD aim to take the further step of cognizing them.

Zero-shot Learning Another relative of OCD is zero-shot learning (ZSL) [21, 32, 33], which directly focuses on recognizing unseen categories. Similar to generalized category discovery [30], a more realistic setting, generalized zero-shot learning (GZSL) [24], where both seen and unseen categories are involved during test is also considered. ZSL/GZSL approaches explicitly leverage sharing side information (*e.g.*, word vectors [28], attributes [14]) across seen and unseen categories to *relate the past and the future*. NCD and OCD aim to implicitly transfer the concept of category from seen to unseen data without relying on prior novel category definition via side information.

Novel Category Discovery Han *et al.* [11] first formalized Novel category discovery (NCD) and tackled this problem via deep transfer clustering that simultaneously learns visual representation and conducts clustering. After that, Han *et al.* [10] proposed a novel framework named AutoNovel with three training stages. In particular, at the last stage, a novel rank statistic technique was proposed to form pairwise pseudo labels for joint training on labelled and unlabeled data. Zhao *et al.* [36] proposed a two-branch network for learning both global and local features and adopted dual-rank statistics. Furthermore, Jia *et al.* [15] applied rank statistics on feature groups and designed a winner-take-all hashing approach. In addition, Fini *et al.* [7] used the Sinkhorn-Knopp algorithm for generating pseudo labels

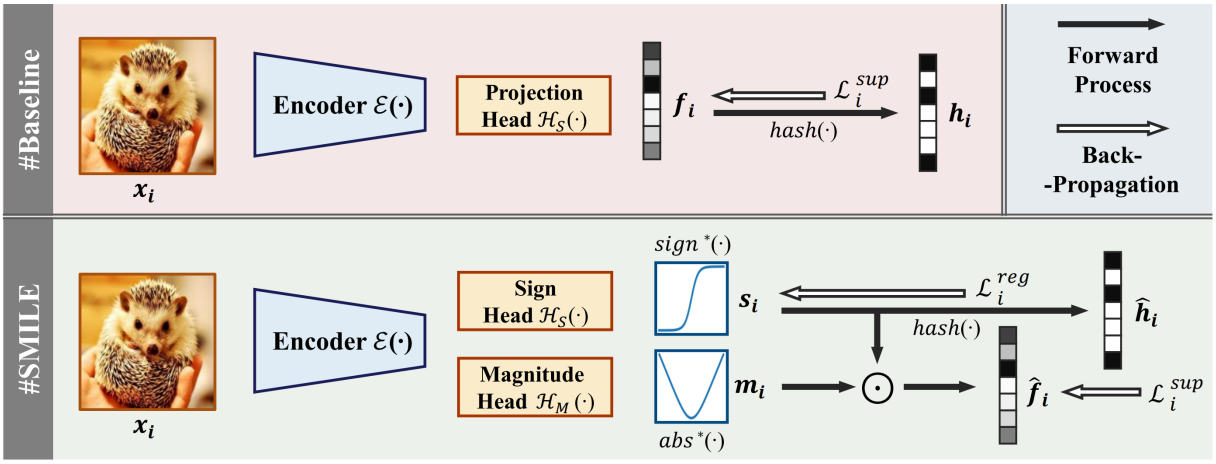


Figure 2. Illustration of the baseline model and the proposed SMILE. The baseline model uses one projection head that generates a feature embedding $\boldsymbol{f}_i$ for both supervised contrastive learning and hashing. In contrast, the proposed SMILE adopts a two-branch structure that separately infers the sign and magnitude information of the feature embedding for supervised contrastive learning. And only outputs from the sign head are hashed to form the category descriptor $\boldsymbol{h}_i$.

and then trained the model with both labelled and unlabeled samples with a unified objective. Recently, Vaze *et al.* [30] introduced a less constrained NCD setting where both seen and unseen categories appear during test, and the category number is unknown in advance. However, two limitations still exist in the current setting: (i) transductive learning – the model is limited to inference on a pre-defined query set that must be available during training, and (ii) offline evaluation – the model is incapable of instant feedback and individual results are dependent on the overall query data distribution. Therefore, this paper extends offline NCD by proposing an inductive, online setting named on-the-fly category discovery (OCD).

3. Methodology

3.1. Overview

In this paper, focusing on letting machines learn to cognize unseen categories, we put forward the problem of on-the-fly category discovery (OCD). Learning is conducted on a closed labelled set, and then known and novel categories should be recognised on-the-fly.

To start with, we define the data structure of OCD as follows. The full data D consists of a support set D_S for training and a query set D_Q for testing. We have $D_S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in X_S \times Y_S$ and $D_Q = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^M \in X_Q \times Y_Q$, where x_i denotes samples within the dataset and y_i is the corresponding labels. Note that, following the setting of generalized category discovery (GCD) [30], the query set D_Q contains both seen categories and unseen categories, *i.e.*, $Y_S \in Y_Q$. The differences between OCD and GCD lie in: (i) only the support set D_S is used for model training, *i.e.*, inductive learning, and (ii) samples in the query set D_Q are individually inferred, *i.e.*, instant feedback.

The methodology is organized as follows: in Section 3.2, we first introduce a hash coding-based strong baseline for OCD, and in Section 3.3, we further proposed a simple yet effective **S**ign-**M**agnitude d**I**sentang**LE**ment (SMILE) architecture that encourages category-consistent hash coding.

3.2. Hash Coding-based Baseline Model

3.2.1 Expandable Prediction Space

Conventional visual recognition models can be formalized as two coupled components: an encoder $\mathcal{E}(\cdot)$ for feature extraction and a classifier $\mathcal{C}(\cdot)$ that projects extracted features into the prediction space. Given a dataset $D \in X \times Y$, for any sample $\boldsymbol{x}_i \in D$, the model output can be written as

$$\hat{y}_i = argmax(\mathcal{C}(\mathcal{E}(\boldsymbol{x}_i))) \in [1, |Y|], \tag{1}$$

where $argmax(\cdot)$ returns the category index with the highest probability. However, this architecture restricts the prediction space to a closed set Y and cannot handle sample with novel categories, *e.g.*, $(\boldsymbol{x}_j, y_j)$ with $y_j \notin Y$.

Existing NCD models, can be roughly divided into two parts: an encoder $\mathcal{E}(\cdot)$ for feature extraction and a projection head $\mathcal{H}(\cdot)$ that projects extracted feature into a discriminative embedding space. Then, labels can be allocated to the query samples via clustering techniques. Such a decision process frees the model from in-category prediction. Thus we follow this road to construct our baseline model.

For *inductive* training, we only use the labelled support data D_S for learning $\mathcal{H}$ and $\mathcal{E}$. As such we only apply supervised contrastive learning [16] here. Letting $\boldsymbol{f}_i = \mathcal{H}(\mathcal{E}(\boldsymbol{x}_i))$ be the feature embedding for $\boldsymbol{x}_i$, the optimization goal can be formulated as

$$\mathcal{L}_i^{sup} = -\frac{1}{|P_i|} \sum_{p \in P_i} log \frac{exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_p)}{\sum_{j=1}^{|B|} \mathbb{1}_{[j \neq i]} exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_j)}, \tag{2}$$

where B is the multi-view-augmented batch that contains $\boldsymbol{x}_i$, P_i denotes the indices of other positive samples in B.

3.2.2 Instant Inference via Hash Coding

To achieve instant feedback, the widely used clustering techniques in NCD are no longer practical. Therefore, to form decision boundaries that can recognize both seen and unseen categories, an intuitive choice is setting a threshold as the minimum inter-category distance. However, such metric-based decision-making is still impractical for inference on individual samples, and does not support cognizing novel categories for which cannot form category prototypes.

To support online recognition of known categories and the ability to cognize unknown categories, our novel solution is to regularize the embedding space to a structural hash space where decision boundaries naturally exist – simply letting each category correspond to a specific hash code. For any sample x_i, we can obtain its hash identifier by

$$\boldsymbol{h}_i = hash(\mathcal{H}(\mathcal{E}(\boldsymbol{x}_i))) \in \{0,1\}^L, \tag{3}$$

where L is the length of hash codes and $hash(\cdot)$ is, for an arbitrary vector $\boldsymbol{a} = [a_1, \ldots, a_l, \ldots, a_L]$,

$$hash(\boldsymbol{a}) = [a_1^*, \ldots, a_l^*, \ldots, a_L^*],\ a_l^* = \begin{cases} 1 & a_l \geq 0 \\ 0 & a_l < 0 \end{cases}. \tag{4}$$

During test, we regard samples with the same hash code as a cluster. Unlike existing clustering based solutions, this algorithm can be applied on-the-fly as the inference for each $\boldsymbol{x}_i$ does not depend on the overall query data distribution. We evaluate our hash-based clustering with standard NCD metrics (see Section 4.1 for details). It is worth noting that this hash-based decision space has a maximum number of categories 2^L. Still, it does make the model expandable to unseen categories. The effect of code length L is discussed in Section 4.3.

3.3. Sign-Magnitude Disentanglement

Although a hash coding-based network can make inference on-the-fly, a limitation remains – the hash coding is sensitive to intra-category variance, resulting in a large number of non-existent categories being predicted. The reason behind this is relatively straightforward – assuming each dimension of features/hash codes stands for a specific high-level semantic (*e.g.*, the sharp mouths of hedgehogs), it may not be consistently expressed for all samples (*e.g.*, a curled hedgehog hiding its mouth). Ideal category hash codes should represent the category-level semantics; however, features may respond to instance-level semantic expressions in practice.

To address this, we introduce **S**ign-**M**agnitude d**I**sentang**LE**ment (SMILE). After we hash the feature embeddings according to Equation 4, only the sign information is kept for the final decision, while the magnitude information is discarded. By decoupling the feature embedding into the sign and the magnitude part, then ideally: (i) the sign part represents the inherent semantics of the object's latent category, while (ii) the magnitude part indicates the instance-level expression of each semantic, and (iii) the prediction of signs is independent of the magnitudes. Based on this thinking, we propose a novel sign-magnitude disentanglement that separately infers signs and magnitudes of the feature embeddings to improve category consistent hash coding.

Specifically, we introduce a two-branch architecture where the projection head $\mathcal{H}(\cdot)$ in the baseline model is replaced with a sign-head $\mathcal{H}_S(\cdot)$ and a magnitude-head $\mathcal{H}_M(\cdot)$ in parallel. For the sign-head, we equip it with a sign activation function to discard magnitude information, *i.e.*, $\boldsymbol{s}_i = sign(\mathcal{H}_S(\mathcal{E}(\boldsymbol{x}_i)))$. As for the magnitude-head, we take the output's absolute value to discard sign information, *i.e.*, $\boldsymbol{m}_i = abs(\mathcal{H}_M(\mathcal{E}(\boldsymbol{x}_i)))$. Afterwards, we take their multiplication $\hat{\boldsymbol{f}}_i = \boldsymbol{s}_i \otimes \boldsymbol{m}_i$ as the final feature embedding, where $\otimes$ denotes element-wise product.

During training, we use $\hat{\boldsymbol{f}}_i$ to calculate the supervised contrastive loss $\mathcal{L}_i^{sup}$. And only the output from the sign-head is used for inference, *i.e.*, the hash-based descriptor is $\hat{\boldsymbol{h}}_i = hash(\mathcal{H}_S(\mathcal{E}(\boldsymbol{x}_i)))$. In practice, we adopt the smoothed version of $sign(\cdot)$ and $abs(\cdot)$ function for easier gradient propagation as

$$sign^*(\boldsymbol{a}) \approx (e^{\boldsymbol{a}\times\tau} - e^{-\boldsymbol{a}\times\tau}) \oslash (e^{\boldsymbol{a}\times\tau} + e^{-\boldsymbol{a}\times\tau}), \tag{5}$$

$$abs^*(\boldsymbol{a}) \approx \boldsymbol{a} \odot (e^{\boldsymbol{a}\times\tau} + e^{-\boldsymbol{a}\times\tau}) \oslash (e^{\boldsymbol{a}\times\tau} - e^{-\boldsymbol{a}\times\tau}), \tag{6}$$

where $\boldsymbol{a}$ is an arbitrary vector, τ is a hyper-parameter that controls the smoothness of two functions, $\oslash$ and $\odot$ denote the element-wise division and multiplication, respectively. Note that when $\mathcal{H}_S$ and $\mathcal{H}_M$ share the same weights, $\hat{\boldsymbol{f}}_i$ could degrade to $\boldsymbol{f}_i$.

In addition, as we use the smoothed sign function $sign^*(\cdot)$, we introduce an magnitude regularization (inverse-L1 regularization) term $\mathcal{L}_i^{reg} = -|\hat{\boldsymbol{h}}_i|$ to encourage $|\hat{\boldsymbol{h}}_i|$ to be close to 1 or -1, which also implicitly encourages the sign-head to be less sensitive to intra-category variance. The total optimization goal is then

$$\mathcal{L}_i^{total} = \mathcal{L}_i^{sup} + \alpha\mathcal{L}_i^{reg}, \tag{7}$$

where α is a hyper-parameter that balances the effects of two loss functions. The ablation studies about the effectiveness of each component, the effects of different hyper-parameters τ, and different α can be found in Section 4.3.

4. Experiment

4.1. Experiment Setup

Datasets Following the setup of GCD [30], we adopt six datasets in our experiments, including three coarse-grained

	CIFAR10	CIFAR100	ImageNet-100	CUB	CAR	HERB19
$\lvert Y_S\rvert$	5	80	50	100	98	341
$\lvert Y_Q\rvert$	10	100	100	200	196	683
$\lvert D_S\rvert$	12.5K	20.0K	31.9K	1.5K	2.0K	8.9K
$\lvert D_Q\rvert$	37.5K	30.0K	95.3K	4.5K	6.1K	25.4K

Table 1. Statistics of datasets used in our experiments. Number of categories $|Y_S|$ & $|Y_Q|$ and number of samples $|D_S|$ & $|D_Q|$.

classification datasets: CIFAR10 [19], CIFAR100 [19], Imagenet-100 [25], and three fine-grained classification datasets: CUB-200-2011 [31], Stanford Cars [18], and Herbarium19 [29]. Note that ImageNet-100 refers to a subset of ImageNet with 100 categories randomly sampled. The categories of each dataset are split into subsets of seen and unseen categories. And 50% samples belonging to the seen categories are used to form the support set D_S for training, and the rest form the query set D_Q for testing. The statistics of split datasets can be found in Table 1.

Models for Comparison Since the overall distribution of the query set is unavailable, many previous approaches are impractical under our instant inference setting, *e.g.,* clustering-based approaches [30], pseudo-labelling via extended classifier [34], and pseudo-label allocation via Sinkhorn-Knopp algorithm [7]. To demonstrate the superiority of SMILE, we modify the following methods to meet our setting for comparison:

(1) Baseline: Our simple hashing framework without SMILE.

(2) Sequential Leader Clustering (SLC) [12]: SLC is a classical clustering technique for scenarios where data come in sequence. It owns the advantage that the cluster number does not need to be defined in advance. We replace the hash coding-based decision-making on our baseline with SLC as a fair competitor.

(3) Meta-learning for Domain Generalization (MLDG) [20]: Unlike the conventional NCD task that leverages both the support and the query set for discriminative feature learning, the proposed OCD is more like a generalization problem – learning from seen categories and then generalizing to unseen categories. Therefore, in this paper, we adopt the model agnostic domain generalization algorithm MLDG [20] as a competitor. In particular, samples from different categories are split into meta-train and meta-test domains at each training iteration, and then we apply MLDG to our baseline model.

(4) Ranking Statistices (RankStat) [10]: Ranking statistics are employed by AutoNovel [10] to measure sample relationships – belonging to the same category or not. Specifically, it leverages the top-3 indices of feature embeddings as category descriptors, and samples with the same top-3 indices are likelier to belong to the same category. This meets our OCD setting and becomes a strong competitor to our hash-based category descriptor. We keep the self-supervised and fully-supervised learning stages for comparison. We discard the third stage of joint optimization with pseudo-labelled samples since the query set does not participate in model training in the OCD setting.

(5) Winner-take-all (WTA) [15]: Considering the ranking statistics may excessively focus on salient features and overlook the holistic structure information. In [15], the authors proposed winner-take-all hash as an alternative. In particular, instead of using the global order of feature embeddings, WTA takes the indices of maximum values within feature groups. Here we take the whole WTA codes as descriptors for instant inference.

Implementation Details For fair comparison, we take DINO [2] pre-trained ViT-B-16 [6] as the backbone network (*i.e.*, the encoder $\mathcal{E}(\cdot)$) for all models. We unified the fully-supervised learning scheme for all methods as supervised contrastive learning [16], as it performs better than linear classifier used by some previous SOTA approaches (*e.g.*, RankStat [10]). We fine-tune the final block of ViT-B-16 for all methods.

All methods are trained for 50 epochs on coarse-grained datasets and 100 epochs on fine-grained datasets. For optimization, we use SGD with the momentum of 0.9, the start learning rate of 0.01, and the cosine learning rate decay schedule [22]. We take batch size $|B| = 128$ for all methods, and hyper-parameter $\alpha = 3$ and $\tau = 1$ for SMILE (related experimental results can be found in Section 4.3).

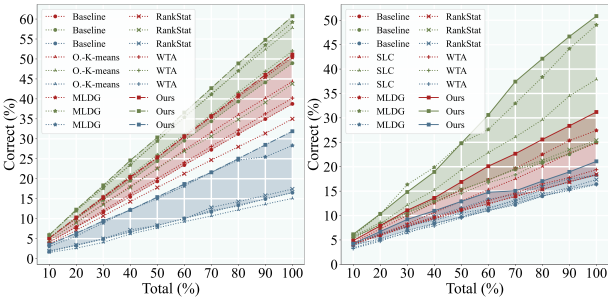
For the hash code length, we take $L = 12$ for all hash coding-based models. It is worth noting that $L = 12$ does not consistently yields the best results on different datasets, but we keep this value the same since we do not know the novel category number in advance. The effect of the code length L is discussed in Section 4.3. Besides, to ensure the prediction spaces of similar size, for RankStat, we set the embedding dimension to 32 with top-3 indices being focused; for WTA, we use 48-dimension embeddings divided into 3 groups[1].

Evaluation Protocols This paper adopts two protocols for evaluation termed **Greedy-Hungarian** and **Strict-Hungarian** for comprehensive comparisons. The two protocols are used by Fini *et al.* [7] and Vaze *et al.* [30], respectively, and their difference is clearly illustrated in [30]. During the testing phase, we regard samples with the same category descriptor as a cluster. All clusters are sorted according to their sizes. We only keep the top-$|Y_Q|$ clusters, and the rest clusters will be treated as misclassified. Afterwards, for **Greedy-Hungarian**, samples are first divided into the "New" and "Old" sub-set by their ground-truth labels, and then we calculate the accuracy of each sub-set separately. Thus it provides an independent perspective of model performance on the "New" and "Old" sub-set. As for

[1]The prediction space size of hash coding, RankStat, and WTA are $2^{12} = 4096$, $C_{32}^{3} = 4960$, and $16^3 = 4096$, respectively.

	Method	CIFAR10 (%)			CIFAR100 (%)			ImageNet-100 (%)			CUB-200-2011 (%)			Stanford Cars (%)			HERB19 (%)		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Greedy-Hungarian	Baseline	64.82	94.47	49.99	40.36	50.94	19.20	34.38	76.52	13.20	22.63	29.58	19.16	16.12	23.86	12.37	14.72	22.88	10.33
	SLC [12]	65.92	96.49	50.86	46.90	62.05	16.57	34.18	86.55	7.14	30.20	46.50	22.06	14.39	23.94	9.78	15.95	28.54	9.14
	MLDG [20]	71.58	97.50	58.62	58.42	68.99	37.27	33.59	74.40	13.08	34.21	57.91	22.36	27.96	49.13	17.74	23.78	39.47	15.33
	RankStat [10]	56.51	81.07	44.23	36.87	45.67	19.27	33.07	74.19	12.40	22.70	29.95	19.08	16.23	23.29	12.82	15.03	22.25	11.15
	WTA [15]	65.38	87.98	54.08	44.09	55.52	21.24	33.12	75.82	11.66	23.78	30.51	20.42	18.34	26.29	14.51	16.18	23.21	12.39
	SMILE (Ours)	**78.17**	**99.27**	**67.62**	**61.31**	**70.71**	**42.48**	**39.91**	**87.07**	**16.22**	**41.11**	**67.65**	**27.84**	**33.35**	**58.41**	**21.25**	**28.36**	**45.56**	**19.11**
Strict-Hungarian	Baseline	43.53	56.21	37.30	37.82	48.75	15.96	31.49	72.92	10.68	21.09	26.19	18.54	15.43	23.04	11.74	13.95	22.19	9.51
	SLC [12]	41.54	**58.29**	33.29	44.36	58.98	15.10	32.92	**86.55**	5.22	28.60	43.96	20.92	14.01	23.04	9.65	14.92	27.44	8.14
	MLDG [20]	44.14	38.47	46.98	50.60	60.98	29.83	30.63	72.30	9.69	29.52	48.37	20.09	23.96	41.63	15.42	20.84	36.67	12.33
	RankStat [10]	42.14	49.26	38.59	35.00	44.01	16.98	31.06	73.30	9.83	21.19	26.85	18.35	14.78	19.94	12.29	13.81	20.63	10.15
	WTA [15]	43.12	34.52	47.42	40.83	52.89	16.72	30.84	72.92	9.68	21.93	26.93	19.43	17.09	24.37	13.59	14.62	21.21	11.07
	SMILE (Ours)	**49.86**	39.86	**54.86**	**51.59**	**61.55**	**31.69**	**33.78**	74.22	**13.45**	**32.24**	**50.89**	**22.91**	**26.15**	**46.65**	**16.25**	**22.90**	**39.29**	**14.09**

Table 2. Comparison with other SOTA methods. The best results are marked in **bold**, and the second best results are marked by underline.



(a) CIFAR100 (b) CUB-200-2011

Figure 3. Illustration of correct cognized samples *versus* total samples as data are input in stream form. Lines of red, green, and blue represent the "All", "Old", and "New" accuracy, respectively. And zones with the corresponding colours indicate the performance margin between SMILE and the baseline.

the **Strict-Hungarian**, the accuracy of the whole query set is first calculated, which prevents the situation that a cluster is repeatedly used by the "New" and "Old" categories. The accuracy calculation via Hungarian algorithm can be formulated as

$$Acc = \max_{s(\cdot)\in\mathcal{S}(Y_Q)} \frac{1}{|D_Q|} \sum_{i=1}^{|D_Q|} \mathbb{1}[\hat{y}_i = s(y_i)], \qquad (8)$$

where y_i is the ground truth label, $\hat{y}_i$ is the predicted label decided by cluster indices, and $\mathcal{S}(Y_Q)$ is the set of all permutations of ground truth labels.

4.2. Comparison with SOTA Methods

We conduct comparison experiments with the aforementioned competitors on all 6 datasets. The experimental results are reported in Table 2. Compared with the baseline model, the proposed SMILE obtained a significant average improvement of 8.54%, which demonstrate the effectiveness of our sign-magnitude disentanglement idea. Although RankStat and WTA can also combat sample noises by focusing on only top-ranking indices, they have not consistently surpassed the baseline model, which suggests that simply ignoring low response channels cannot eliminate the effect of intra-category variance.

On the contrary, the classical domain generalization algorithm MLDG obtained excellent results – the second-best overall performance. This verifies our judgement that OCD, in a sense, can be regarded as a generalization problem. Instead of domain generalization that transfers knowledge to unknown domains, OCD requires the cognition ability can be transferred to unknown categories. Therefore, we might state that a possible direction for better solving OCD tasks can be generalizable feature learning.

In addition, in Figure 3, we illustrate the real-time performance on CIFAR100 and CUB-200-2011 as data are input in stream form. The real-time performance gains are also illustrated in the figure. We can find the model performance is quite stable over different amounts of data – the relation between the total input data percentage and the correct cognized percentage is almost linear. It is intuitive as SMILE conducts inductive inference and does not leverage the input data to improve model performance. And, of course, an unsupervised incremental learning process with knowledge cumulation could be an interesting future direction.

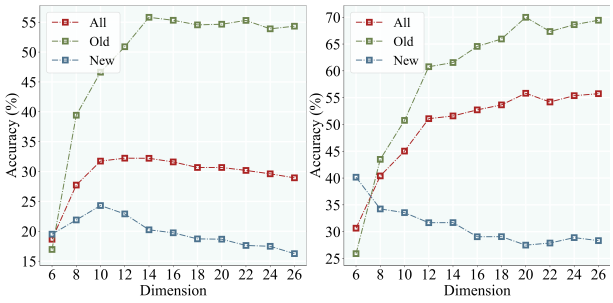
Besides, it is worth noting that, despite the consistent overall trends under the two evaluation protocols, exceptional cases also exist. *E.g.*, SMILE achieves the best result on CIFAR10 under the Greedy-Hungarian protocol but obtains only 39.86% accuracy for the old category under the Strict-Hungarian protocol. The reason behind this is that too few categories in CIFAR10 make the model biased to seen categories – when an old category and a new category are clustered together, the Hungarian algorithm will allocate this cluster to the new category for the best overall accuracy since there are more samples from new categories in the test set. Therefore, we adopt both protocols to understand these exceptional cases better.

4.3. Ablation Studies

In this section, we conduct ablation studies on CIFAR100 and CUB-200-2011, respectively, the typical

S.-M. Disen.	M. Reg.	CIFAR100 (%)			CUB-200-2011 (%)		
		All	Old	New	All	Old	New
✗	✗	37.82	48.75	15.96	21.09	26.19	18.54
✓	✗	43.94	51.43	28.97	24.93	29.34	22.72
✗	✓	42.05	53.92	18.31	21.48	26.74	18.85
✓	✓	**51.59**	**61.55**	**31.69**	**32.24**	**50.89**	**22.91**

Table 3. Ablation study on the sign-magnitude disentanglement and magitude regularization. The best results are marked in **bold**. S.-M. Disen. and M. Reg. stand for sign-magnitude disentanglement and magnitude regularization, respectively.



(a) CIFAR100 (b) CUB-200-2011

Figure 4. Results with different hash code dimension L.

coarse-grained and fine-grained datasets.

Effectiveness of Each Components: The proposed SMILE can be divided into two components: the **sign-magnitude disentanglement** architecture and the **magnitude regularization**. In this sub-section, we investigate the effectiveness of each component and the benefit of their cooperation. When the magnitude regularization works alone, we directly apply it to the projection head of our baseline model. As shown in Table 3, each component and their combination consistently boost model performance. And the interesting point is that when they work alone, an average improvement of 9.96% and 2.31% is respectively obtained; when they work together, the average gain is 12.46%, which is even more significant than the sum of two individual improvements. It indicates that their working mechanisms are highly coupled – the magnitude regularization encourages the sign head to ignore magnitude information, and the sign-magnitude disentanglement ensures magnitude information is kept for contrastive learning.

Effects of the Hash Code Length L: The code length L is the most critical factor for hash coding-based models. Under the setting of OCD, it directly decides the size of prediction space (equal to 2^L). An ample prediction space means a better capability to handle large-amount categories and finer representations with more details. On the contrary, a smaller L leads to more robust feature representations against intra-category variance.

Here we experiment with L varying from 6 to 20, as shown in Figure 4. We can observe that the best performance is obtained with different dimensions for seen and unseen categories, even on the same dataset. The model works better on unseen categories with a smaller L since more robust feature embeddings are required for a better generalization ability. In addition, we notice the overall performance benefits more from large L on the coarse-grained dataset, which may be because coarse-grained recognition is less sensitive to the effect of the intra-category variance than fine-grained recognition.

In other experiments, we take $L = 12$ for all datasets instead of selecting the optimal choice. This is because the code length L, which decides the size of prediction space, directly relates to the number of unseen categories we have no prior knowledge of in the OCD setting. And designing an optimal code length estimation technique that works on the support set may be a meaningful future direction.

Effects of the Hyper-parameter α: Hyper-parameter α controls the contribution of the magnitude regularization loss. As shown in Table 4, we experiment with α within $\{1, 2, 3, 4, 5\}$. And the model performance is not that sensitive with the amount of α – similar results are obtained with α from 1 to 3. We take $\alpha = 3$ for all experiments.

Effects of the Hyper-parameter τ: Hyper-parameter τ controls the smoothness of the approximate $sign(\cdot)$ and $abs(\cdot)$ function. We take the same τ for both two functions to ensure their multiplication is still a linear function. A higher τ results in sharper functions, which means better disentanglement. Conversely, smoother functions with lower τ facilitate gradient propagation. We experiment with $\tau \in \{0.2, 0.5, 1, 2, 5\}$, as shown in Table 5. We take $\tau = 1$ for all other experiments.

4.4. Discussion

What defines the concept of "category"? The reader may raise the question that different people may hold various understandings about the word "category", *e.g.*, the same bird can be "Flamingo", "Phoenicopteridae", and "Phoenicopteriformes" to people of different expertise about birds. Then what defines the expertise level of OCD models? To this end, we experiment on CUB-200-2011, and Stanford Cars with the support set annotated at different granularity levels[2].

Experimental results are reported in Table 6, and three expertise levels are expressed by "*Novice*", "*Erudite*", and "*Expert*". The "*Expert*" level annotations are what we generally use in fine-grained recognition, and we still evaluate the model at this level on the query set. In addition, we also report the number of categories we discovered (*i.e.*, the number of different hash codes). According to the results, we can come to a very intuitive conclusion – the expertise level of the OCD model is defined by the annotation level of

[2]We adopt the hierarchical labels in [4]. The CUB-200-2011 dataset is annotated by three-level labels of "Order", "Family", and "Species". Stanford Cars is annotated into two-level labels by "Type" and "Model".

Hyper-parameter α	CIFAR100 (%)			CUB-200-2011 (%)		
	All	Old	New	All	Old	New
1	49.96	60.64	28.61	31.61	50.90	21.96
2	51.24	**61.67**	30.38	32.03	**51.90**	22.09
3	**51.59**	61.55	31.69	**32.24**	50.89	**22.91**
4	45.80	52.32	**32.76**	30.14	45.19	22.50
5	42.69	53.24	21.60	28.60	42.29	21.76

Table 4. Results with different Hyper-parameter α. The best results are marked in **bold**.

Hyper-parameter τ	CIFAR100 (%)			CUB-200-2011 (%)		
	All	Old	New	All	Old	New
0.2	50.81	**64.56**	23.31	31.43	**58.11**	18.08
0.5	50.81	62.53	27.38	31.90	55.50	20.09
1	**51.59**	61.55	**31.69**	**32.24**	50.89	**22.91**
2	49.16	60.18	27.12	26.71	35.09	22.52
5	46.31	58.75	21.42	19.35	20.35	18.85

Table 5. Results with different Hyper-parameter τ. The best results are marked in **bold**.

Expertise Level	CUB-200-2011 (%)				Stanford Cars (%)			
	All	Old	New	Num.	All	Old	New	Num.
Novice	10.34	12.21	9.41	81	10.51	11.74	9.92	51
Erudite	16.30	16.94	15.98	308	N/A	N/A	N/A	N/A
Expert	**32.24**	**50.89**	**22.91**	436	**26.15**	**46.65**	**16.25**	343

Table 6. Results when the support set is annotated with different granularity levels. The best results are marked in **bold**. Num. stands for the inferred category number.

Transfer Direction	Overall Acc (%)	Transfer Direction	Overall Acc (%)	Transfer Direction	Overall Acc (%)
$P_1 \rightarrow G$	7.61	$P_2 \rightarrow P_1$	10.53	$P_1 \rightarrow P_2$	7.71
$P_2 \rightarrow G$	8.81	$G \rightarrow P_1$	16.77	$G \rightarrow P_2$	13.92
$G \rightarrow G$	33.78	$P_1 \rightarrow P_1$	32.24	$P_2 \rightarrow P_2$	26.15

Table 7. Results for different knowledge transfer scenarios with one general field and two professional fields, ImageNet-100 (G), CUB-200-2011 (P_1), and Stanford Cars (P_2).

the support set. In particular, the model performance significantly drops as the annotation level degrades because the model trained on coarse-grained annotations cannot distinguish fine-grained categories. Besides, the number of discovered categories also decreases, *i.e.*, many fine-grained categories are merged into the same category, which indicates the degradation of the model's expertise level.

Can we achieve cross-domain discovery? Similar to the well-analysed NCD problem, the OCD relies on the learning and transferring of shared semantics from support to query. A task with few shared semantics between the support and query set may be unsolvable. To test the limitations of OCD, we evaluate with three datasets: CUB-200-2011 (P_1) and Stanford Cars (P_2) represent two separate professional fields, and ImageNet-100 (G) stands for a general field. We synthesise four different knowledge transfer scenarios: (i) transferring within the professional fields ($P_1 \rightarrow P_1$ and $P_2 \rightarrow P_2$), (ii) transferring across professional fields ($P_1 \rightarrow P_2$ and $P_2 \rightarrow P_1$), (iii) transferring from professional fields to general fields ($P_1 \rightarrow G$ and $P_2 \rightarrow G$), and (iv) transferring from general fields to professional fields ($G \rightarrow P_1$ and $G \rightarrow P_2$). Note that we also report results of "$G \rightarrow G$" for comparison; however, as we assume the general field is unique, there is no concept about "within the general field" or "across the general field".

Experimental results are shown in Table 7. We can observe that the model performance significantly degrades when we transfer knowledge across different professional fields due to the limited high-level semantics they share, *e.g.*, it is hard to find common ground between birds and cars. As for transferring knowledge from a general field to professional fields, the results are slightly better but still not satisfactory – although a general field stands a good chance of covering semantics of professional fields, models trained for coarse-grained recognition are incapable of cognizing fine-grained categories [4]. Similarly, transferring from professional to general fields also leads to worse performance because they might only cover a small subset of high-level semantics in the general fields.

These observations suggest that cognizing unseen categories is not a panacea. It echos the statement in [5] that novel category discovery tasks should be designed following a *sampling in causality* protocol, *i.e.*, the seen and unseen categories should be collected in the same way; otherwise, the task may tend to be theoretically unsolvable.

5. Conclusion

In this paper, we put forward a new and highly practical visual recognition problem termed on-the-fly category discovery (OCD). Meanwhile, an intuitive hash coding-based baseline model is designed as a solution for OCD, which we further improve with a simple yet effective sign-magnitude disentanglement architecture. Comprehensive experiments on 6 popular datasets verify the effectiveness of the proposed method. Additionally, we also discuss the limitation of OCD via experiments with multi-granularity support set annotations and cross-domain support sets.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (NSFC) No. U19B2036, 62106022, 62225601, in part by Beijing Natural Science Foundation Project No. Z200002, in part by scholarships from China Scholarship Council (CSC) under Grant CSC No. 202206470055, 202006470036, in part by BUPT Excellent Ph.D. Students Foundation No. CX2022152, CX2020105, in part by the Program for Youth Innovative Research Team of BUPT No. 2023QNTD02, and in part by the Supported by High-performance Computing Platform of BUPT.

References

[1] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, 2015. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5

[3] Hakan Cevikalp and Bill Triggs. Polyhedral conic classifiers for visual object detection and classification. In *CVPR*, 2017. 2

[4] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your" flamingo" is my" bird": Fine-grained, or not. In *CVPR*, 2021. 7, 8

[5] Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, Gang Niu, Mingyuan Zhou, and Masashi Sugiyama. Meta discovery: Learning to discover novel classes given very limited data. In *ICLR*, 2021. 8

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5

[7] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *ICCV*, 2021. 1, 2, 5

[8] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017. 2

[9] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 1, 2

[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. 1, 2, 5, 6

[11] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019. 1, 2

[12] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975. 5, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 1

[14] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, 2020. 2

[15] 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. 1, 2, 5, 6

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 2020. 2, 3, 5

[17] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *ICCV*, 2021. 2

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 5

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

[20] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 5, 6

[21] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *CVPR*, 2022. 2

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[23] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018. 2

[24] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5

[26] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 2

[27] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2014. 2

[28] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 2013. 2

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

[30] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 2, 3, 4, 5

[31] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5

[32] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–37, 2019. 2

[33] Kun Wei, Cheng Deng, Xu Yang, et al. Lifelong zero-shot learning. In *IJCAI*, 2020. 2

[34] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Self-labeling framework for novel category discovery over domains. In *AAAI*, 2022. 5

[35] He Zhang and Vishal M Patel. Sparse representation-based open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1690–1696, 2016. 2

[36] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *NeurIPS*, 2021. 2

[37] 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. 1