Novel Class Discovery without Forgetting

K J Joseph[1,2], Sujoy Paul[1], Gaurav Aggarwal[1], Soma Biswas[3], Piyush Rai[1,4], Kai Han[1,5], and Vineeth N Balasubramanian[2]

[1] Google Research
[2] Indian Institute of Technology Hyderabad
[3] Indian Institute of Science, Bangalore
[4] Indian Institute of Technology Kanpur
[5] The University of Hong Kong
{cs17m18p100001, vineethnb}@iith.ac.in, somabiswas@iisc.ac.in,
{sujoyp, gauravaggarwal}@google.com, piyush@cse.iitk.ac.in,
kaihanx@hku.hk

Abstract. Humans possess an innate ability to identify and differentiate instances that they are not familiar with, by leveraging and adapting the knowledge that they have acquired so far. Importantly, they achieve this without deteriorating the performance on their earlier learning. Inspired by this, we identify and formulate a new, pragmatic problem setting of *NCDwF: Novel Class Discovery without Forgetting*, which tasks a machine learning model to incrementally discover novel categories of instances from unlabeled data, while maintaining its performance on the previously seen categories. We propose 1) a method to generate pseudo-latent representations which act as a proxy for (no longer available) labeled data, thereby alleviating forgetting, 2) a mutual-information based regularizer which enhances unsupervised discovery of novel classes, and 3) a simple Known Class Identifier which aids generalized inference when the testing data contains instances form both seen and unseen categories. We introduce experimental protocols based on CIFAR-10, CIFAR-100 and ImageNet-1000 to measure the trade-off between knowledge retention and novel class discovery. Our extensive evaluations reveal that existing models catastrophically forget previously seen categories while identifying novel categories, while our method is able to effectively balance between the competing objectives. We hope our work will attract further research into this newly identified pragmatic problem setting.

Keywords: Novel Class Discovery, Catastrophic Forgetting, Generalized Inference, Regularizers, Pseudo-latent Generation and Replay.

1 Introduction

Over the last decade, deep learning algorithms have achieved remarkable performances on multiple computer vision tasks [15,37,50,7,45], even outperforming humans on many of them. These algorithms are specialised to work well in their strictly designed problem setting, but are brittle when the assumptions are relaxed. We closely analyse one such setting here. Current image classification

Fig. 1: Our existing knowledge about birds helps us to easily identify two groups in these images even if we have not seen images of these bird species before. At the same time, unsupervisedly discovering these novel categories does not make us forget about previously seen categories. Motivated by this observation, we propose *NCDwF* setting and a methodology to instill this capability into machines.

models assume availability of training examples of all classes of interest. Once trained and deployed, it recognises instances of classes that it has been taught. An instance outside this set of classes may be wrongly classified into one of the known classes often with high confidence [48,62,46,18]. In contrast, humans can easily identify instances that they do not know, and even differentiate among them. To aid our discussion, let us glance through the set of images in Fig. 1. We naturally concur the following: "These birds are not like anything that we have seen before, but these images do seem to belong to two distinct categories". Importantly, we are able to do this grouping without having access to training images from other objects that we have learnt during our lifetime. Secondly, the ability to do this grouping does not impede us from identifying other kinds of birds that we are already familiar with. Lastly, we achieve this without explicit information that these instances are from novel categories. Motivated by this intrinsic ability of humans, we propose a problem setting, which we refer to as *NCDwF: Novel Class Discovery without Forgetting*.

An NCDwF model learns in phases. In the first phase, the model is supervised to learn a few set of classes. In the subsequent phases, the model should automatically identify instances of novel categories from an unlabeled pool containing instances from a disjoint set of classes. While doing so, model *does not have* access to labeled data from the first phase. At any point in time, the model should classify a test instance to one of the labeled or unlabeled classes, without any task identifying information. Here "task" refers to whether the test instance belongs to a (known) labeled class or a (novel) unlabeled class. We illustrate the problem setting in Fig. 2. After learning about `Bird`, `Dog` and `Elephant` in the first phase, a NCDwF model identifies instances from previously known classes (eg. `Bird`), along with grouping instances of novel categories.

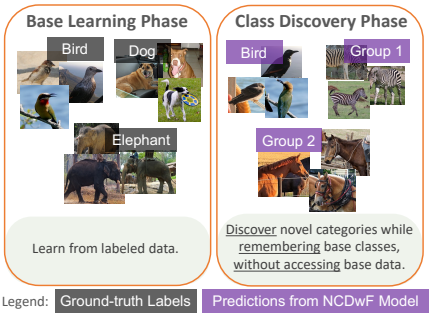

Fig. 2: Summary of NCDwF setting.

The NCDwF setting has wide practical applicability: 1) Consider the recognition component of a robot operating in an open-world. It can be trained in-house with annotated data. Once deployed, it would be of immense value if it can automatically group unknown instances into different groups, along with consistently identifying instances that it has been trained with. 2) Equally interesting would be an online fraud detection system. It can also be trained with a set of known fraud patterns, but it would be hard to speculate emerging frauds. An incremental class discovery model can not only identify novel frauds, but also group them separately, alongside identifying known fraud types, adding immense practical utility. Labeled data that was used to train both these models in-house cannot be accessed while identifying novel instances due to storage and privacy concerns.

NCDwF is closely related to Novel Class Discovery (NCD)[19] but it extends NCD in several key aspects. First, existing NCD methods assume access to both labeled and unlabeled data at training time, which is unlikely to hold for many real applications. Second, at test time, current NCD methods assume access to the "task" information, i.e., the information whether an unlabeled instance is from a labeled class or not. In NCDwF, we relax these assumptions to propose a more pragmatic extension to NCD setting, mirroring real world demands.

Our methodology subtly makes use of the classifier trained on the labeled data to reduce forgetting and improve class discovery. To make up for the lack of labeled examples from previous classes during the unsupervised novel class discovery phase, we identify regions in the latent space by "inverting" the classifier's discriminative information. Additionally, we ensure that these inverted pseudo-latent representations are close to the true class representations as explained in Sec. 3.2. These class specific pseudo-representations can be replayed along with unlabeled data to address forgetting. We note that this method is cheaper than the generative modelling alternatives, and does not require any labeled image to be stored and replayed. In Sec. 3.3, we show that maximizing the mutual information between the labeled logits and the unlabeled logits acts as an effective regularizer to enhance class discovery. The proposed setting calls for a generalized, task-agnostic inference where a test instance may belong to labeled or the unlabeled classes, and such identifying information would be absent during inference. We propose to learn a Known Class Identifier to help us with this discrimination in Sec. 3.4.

To summarize, our key contributions are as follows:

- We propose a pragmatic generalization to the NCD setting called Novel Class Discovery without Forgetting (NCDwF).
- We introduce an effective method which unsupervisedly discovers novel classes, while retaining performance on the labeled classes used to initialize the model.
- We introduce experimental setting and evaluation protocol for the new setting.
- When compared with prominent class-discovery methods [16,19,60] adapted to our proposed setting, our methodology achieves improved class-discovery performance with significantly less forgetting.

Table 1: We summarise related problem settings here. We note that Novel Class Discovery without Forgetting is most pragmatic when compared with the others. ✓, × and − indicates `yes`, `no`, and `not-applicable` respectively. More discussion in Sec. 2.1.

Characteristics (→)	Data from a future step:			
Settings (↓)	can contain disjoint set of classes.	need not have side information.	can make use of a model bootstrapped with labeled data.	can be fully unlabeled.
Semi-supervised Learning	×	−	✓	−
Zero-shot learning	✓	×	✓	✓
One / Few-shot learning	✓	×	✓	×
Clustering	−	−	×	✓
Incremental Learning	−	✓	−	×
NCDwF	✓	✓	✓	✓

2 Related Works

Here, we analyse how NCDwF differs from existing related settings, followed by a survey of research efforts in Incremental Learning and Novel Class Discovery.

2.1 Relation with Existing Settings

We systematically analyse how our proposed setting is related to research efforts in related problem spaces in Tab. 1. NCDwF methods incrementally discover novel category of instances from an unlabeled pool by utilizing the knowledge from a disjoint set of labeled instances. At inference stage, the model should be consistent in classifying instances to any of labeled or unlabeled classes, without any task identifying information. In semi-supervised learning approaches [10,51], the labeled and unlabeled data comes from the same set of classes. Zero-shot learning methods [54,42] require prior knowledge of extra semantic attribute information about the unlabeled classes. Few-shot learning methods [4,57,47] additionally require a few of the unlabeled instances to be labeled. Similar instances are grouped together by clustering algorithms [56,14], but they cannot make use of labeled instances from a disjoint set of classes. Incremental learning methods [41,11,27] learn a single model across tasks, but data for each incremental task is fully annotated. Methods that perform out-of-distribution detection [32,38] and open-set learning [46,18] identify instances significantly different from the training data distribution as novel samples, but do not identify sub-groups within these identified instances automatically. To the best of our knowledge, the proposed setting has minimal assumptions and is most pragmatic, when compared to these settings.

2.2 Incremental Learning

The core focus of incremental learning methods is to alleviate the catastrophic forgetting of neural networks [17,36], when learning a single model across a sequence of tasks. Regularization based methods [31,41,9,53,13,34] ensure that the

parameter adaptations for the new task will be optimal for all the tasks learned so far. Another kind of approach either stores or generates exemplar images for all the tasks introduced to the model so far and replays them while learning a new task [41,34,6,27]. This ensures consistency across all tasks. Dynamically expanding and parameter isolation methods [44,40,39,1,33] form a third class of methods to address forgetting. All these methods require access to labeled instances for all the tasks. In contrast, Novel Class Discovery without Forgetting models identify novel categories from unlabeled data which the model encounters incrementally - without forgetting how to identify instances in the labeled classes which were initially used to bootstrap the model.

2.3 Novel Class Discovery

Earlier methods like MCL [25] and KCL [24] for general transfer learning across domains and tasks meta-learn a binary similarity function from the labeled data and use it to discover classes in the unlabeled data. DTC [20] formalized the problem of Novel Class Discovery and introduced a method based on Deep Embedded Clustering [55] for NCD, by pre-training it on the labeled data followed by learning-based clustering. RS [19] first pretrains the model on the labeled and the unlabeled data with self-supervision and uses ranking statistics to generate pseudo-labels for learning the novel categories. This has been further extended by Zhao and Han [59] to further take local spatial information into account. NCL [60] introduces contrastive learning and OpenMix [61] uses a convex combination of labeled and unlabeled instances to enhance class discovery. UNO [16] learns a unified classifier which identifies labeled and unlabeled instances using ground-truth labels and pseudo-labels respectively. They also introduce a task-agnostic evaluation protocol. Jia *et al.* [26] proposed to leverage contrastive learning with WTA hashing to discover new categories in videos and images.

Existing methods for NCD require access to labeled and unlabeled instances together to discover novel categories, which limits their practical applicability. Most of these methods also assume the unlabeled data only contains instances from new classes or assume the information that whether an unlabeled instance is from new classes is known. The concurrent work by Vaze *et al.* [52] extends NCD to a generalized setting where the unlabeled instances may come from both old and new classes, while still requiring access to labeled and unlabeled instances jointly. In contrast, with Novel Class Discovery without Forgetting, we introduce a staged learning and account for the performance on both labeled and the unlabeled data, without requiring access to the labeled data when learning on unlabeled data to discover new classes. Meanwhile, at test time, we do not assume the unlabeled images are only from new classes nor require to know whether an unlabeled image is from a new class or an old one.

3 Novel Class Discovery without Forgetting

We formally define Novel Class Discovery without Forgetting in Sec. 3.1. NCDwF models should balance between two competing goals: alleviating forgetting of

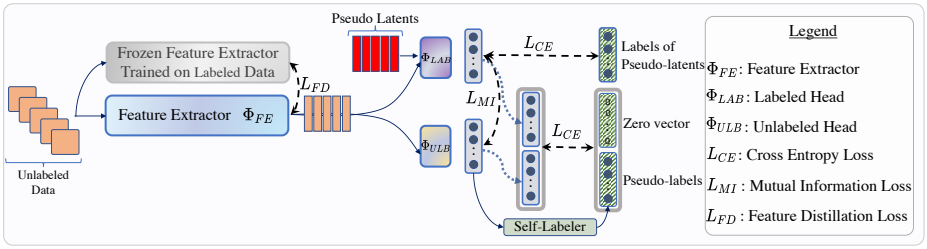

Fig. 3: The figure illustrates how our proposed approach discovers novel categories, while retaining its performance on the labeled data. The network consists of a feature extractor $\mathbf{\Phi}_{FE}$ shared between the labeled head $\mathbf{\Phi}_{LAB}$ and unlabeled head $\mathbf{\Phi}_{ULB}$. We generate pseudo-latents and replay them through the labeled head to reduce its forgetting (Sec. 3.2), and guide the unlabeled head learning through the pseudo-labels and the mutual-information based regularizer (Sec. 3.3).

labeled classes without impairing unsupervised novel class discovery capability. Sec. 3.2 and Sec. 3.3 explain how we achieve these objectives. In Sec. 3.4, we propose Known Class Identifier, which helps with task-agnostic inference.

3.1 Formulation

Given a labeled data pool $D_{lab} = \{(\boldsymbol{x}_i, y_i) \sim P(\mathcal{X}, \mathcal{Y}_{lab})\}$, Novel Class Discovery without Forgetting aims to learn a model $\mathbf{\Psi}$ that would identify novel category of instances from an unlabeled data pool $D_{unlab} = \{(\boldsymbol{x}_i) \sim P(\mathcal{X} \mid \mathcal{Y}_{unlab})\}$, along with recognizing instances from D_{lab}. The label space of D_{lab} and D_{unlab} are disjoint, i.e., $\mathcal{Y}_{lab} \cap \mathcal{Y}_{unlab} = \emptyset$. Further, while discovering novel categories, D_{lab} cannot be accessed. The problem setting naturally induces a multi-stage learning where $\mathbf{\Psi}$ initially learns a representation to identify instances in D_{lab}, which would then be re-purposed to identify novel instances unsupervisedly. The main challenge involved in learning such a $\mathbf{\Psi}$ is to accurately group instances from D_{unlab} into semantically meaningful categories, without degrading its performance on identifying the labeled instances from D_{lab}. Additionally, such a segregation should be done in a generalized fashion, where task identifying information would be absent during inference.

We illustrate the main components of our architecture that help to discover novel categories without forgetting labeled instances in Fig. 3. Without loss of generality, we assume that the model $\mathbf{\Psi}$ consists of a feature extractor $\mathbf{\Phi}_{FE}$, one head for classifying the labeled instances $\mathbf{\Phi}_{LAB}$, and another head for discovering novel categories $\mathbf{\Phi}_{ULB}$. The feature extractor is shared between both heads. Pseudo-latents (shown in red) serve as a proxy for labeled data during category discovery. Pseudo-labels from the self-labeler and the regularization enforced by the mutual-information loss guide the learning of unlabeled head. A frozen model trained only on labeled classes (shown in gray) is also used to regularise the model via feature-distillation loss L_{FD} [22]. We apply an L2 loss between backbone features from the model trained on labeled data $\mathbf{\Phi}^{lab}_{FE}(\boldsymbol{x})$ and current

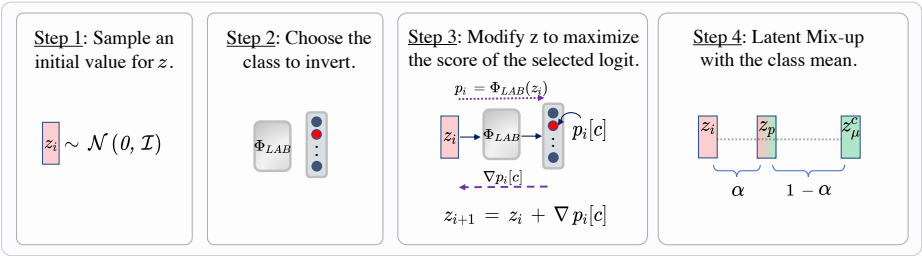

Fig. 4: In the NCDwF setting, labeled data cannot be accessed while discovering novel categories. We propose to generate pseudo-latent representations, which can act as effective proxy for the labeled data representations, by inverting the discriminative information from the trained labeled head as shown in Step 1, 2 and 3. Step 4 helps to induce extra semantic information into these synthesised pseudo-latents.

model $\mathbf{\Phi}_{FE}(\boldsymbol{x})$ as follows: $L_{FD} = \|\mathbf{\Phi}_{FE}(\boldsymbol{x}) - \mathbf{\Phi}^{lab}_{FE}(\boldsymbol{x})\|_2$. Such feature distillation loss has been used in incremental detectors [28] and is simpler than the Less-Forget constraint from LUCIR [23]. The whole model is learned end-to-end, where the feature extractor is free to adapt itself to improve class-discovery, while maintaining its performance of recognizing instances from labeled classes.

3.2 Retaining Performance on Labeled Classes

It would be of immense practical value if a model that is trained in-house with labeled data is able to identify novel category of instances, when deployed in an open world. When the network $\mathbf{\Psi}$ improves its ability to group instances of novel categories from D_{unlab}, it may drastically fail to retain its performance on recognizing the labeled instances, which were learned from D_{lab}, like the well-known *catastrophic forgetting* in lifelong learning [17,36]. This happens as the model cannot be jointly optimised for category discovery and classification of the known instances due to the unavailability of D_{lab}.

We propose a novel methodology that would generate pseudo latent representations, which can act as a proxy for the latent representations of the labeled training data. We make use of the classifier $\mathbf{\Phi}_{LAB}$ that was trained solely on the labeled classes to generate these pseudo-latent representations $\boldsymbol{z}_p$. We explicitly learn these such that it maximally activates a selected class of interest. Fig. 4 summarizes the steps involved to invert the latent knowledge from the classification head. First, we sample $\boldsymbol{z}_i$ from a standard Normal distribution, then we select the specific class c for which we would like to generate the pseudo-latents. Next, we do a gradient ascent on $\boldsymbol{z}_i$ such that the score for the selected class c would be higher for the predicted logit vector $\boldsymbol{p}_i = \mathbf{\Phi}_{LAB}(\boldsymbol{z}_i)$. Importantly, the parameter of $\mathbf{\Phi}_{LAB}$ are frozen, while carrying out the latent inversion $\boldsymbol{z}_{i+1} = \boldsymbol{z}_i + \nabla \boldsymbol{p}_i[c]$, where $\boldsymbol{p}_i = \mathbf{\Phi}_{LAB}(\boldsymbol{z}_i)$. Next, we do *mixup* [58] in latent space between inversed latent $\boldsymbol{z}_i$ and corresponding class mean of labeled training instances $\boldsymbol{z}^c_\mu$. Algo. 1 summarises the steps to generate the latent pseudo-dataset D_{pseudo}. In Lines 4 - 7, we invert the latents of the specific class c, using $\mathbf{\Phi}_{LAB}$.

In Lines 8 and 9, we first select a mixing coefficient α, and do a linear combination of the inverted latent $\boldsymbol{z}_L$ and the class mean $\boldsymbol{z}_\mu^c$. The class means $\boldsymbol{K}$ can be computed and stored after the first phase of learning labeled instances. This mixing operation helps to smoothen the latent space and impart additional semantic information to $\boldsymbol{z}_p$. The labeled pseudo-dataset D_{pseudo} is replayed while learning to identify novel categories, to arrest forgetting in labeled head of $\boldsymbol{\Psi}$.

Algorithm 1 Algorithm GeneratePseudoDataset

Input: Number of labeled classes: M; Number of pseudo-data per class: E; Number of inversion iterations: L; labeled head: $\boldsymbol{\Phi}_{LAB}$; Class Means: $\boldsymbol{K} = \{\boldsymbol{z}_\mu^1, \cdots, \boldsymbol{z}_\mu^M\}$; Parameters of the Beta-distribution: γ, ρ.
Output: Labeled pseudo-dataset: D_{pseudo}.

1: $D_{pseudo} \leftarrow [\,]$
2: **for** c in $(1 \cdots$ M$)$ **do**
3: **for** e in $(1 \cdots$ E$)$ **do**
4: $\boldsymbol{z}_1 \sim \mathcal{N}(0, \boldsymbol{I})$
5: **for** i in $(1 \cdots$ L$)$ **do** ▷ Latent Inversion.
6: $\boldsymbol{p}_i = \boldsymbol{\Phi}_{LAB}(\boldsymbol{z}_i)$
7: $\boldsymbol{z}_{i+1} = \boldsymbol{z}_i + \nabla \boldsymbol{p}_i[c]$
8: $\alpha \leftarrow Beta(\gamma, \rho)$
9: $\boldsymbol{z}_p = \alpha \boldsymbol{z}_L + (1-\alpha)\boldsymbol{z}_\mu^c$ ▷ Latent *mixup* [58] with class mean.
10: $D_{pseudo} \leftarrow D_{pseudo} + (\boldsymbol{z}_p, c)$
11: **return** D_{pseudo}

3.3 Enhancing Class Discovery

Motivated by the success of self-labelling algorithms in self-supervised learning [5,8], Fini *et al.* [16] re-purposes it to automatically generate pseudo labels for the unlabeled data. These labels are used to train the unlabeled head $\boldsymbol{\Phi}_{ULB}$. A key characteristic of such a self-labelling function would be to discourage degenerate solutions. This is explicitly enforced by pseudo-labeling a mini-batch such that the data-points are split uniformly across all the N classes in the unlabeled pool [8,16]. Formally, let $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_B\}$ be the predictions from $\boldsymbol{\Phi}_{ULB}$ for a mini-batch of unlabeled data. Let each mini-batch contains B instances. We seek to find label assignment $\boldsymbol{Q}^* = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_B\}$, such that it respects heterogeneous cluster assignment. This setting can be reduced to solving the following optimal transport problem [5,8]:

$$\boldsymbol{Q}^* = \max_{\boldsymbol{Q} \in \boldsymbol{\mathcal{Q}}} \mathrm{Tr}(\boldsymbol{Q}^\top \boldsymbol{P}) - \sum_{i,j} \boldsymbol{Q}_{ij} \log \boldsymbol{Q}_{ij} \tag{1}$$

where $\boldsymbol{\mathcal{Q}}$ is the transportation polytope defined as $\boldsymbol{\mathcal{Q}} = \{\boldsymbol{Q} \in \mathbb{R}_+^{N \times B} | \boldsymbol{Q}\boldsymbol{1}_{\boldsymbol{B}} = \frac{1}{N}\boldsymbol{1}_N, \boldsymbol{Q}^\top \boldsymbol{1}_N = \frac{1}{B}\boldsymbol{1}_B\}$. An iterative Sinkhorn-Knopp algorithm [12] can be used

to solve Eqn. 1 to find the optimal pseudo-label $\boldsymbol{Q}^*$. The assumption that each mini-batch will be partitioned into all unlabeled classes is fallible. This would lead to noisy pseudo-labels that are not semantically grounded. We are motivated by the observation that labeled head confidently predicts unlabeled data-points into one of the semantically related known categories. For instance, a `motorcycle` gets misclassified into semantically related `bicycle`, and not into other classes that are completely unrelated (more examples in Sec. 5.2). We propose a method which complements the learning via the self-labeled pseudo label by using the semantic information that is available for free within the labeled head.

In the first stage of NCDwF, instances from the labeled data pool D_{lab} would be introduced to the model $\boldsymbol{\Psi}$. We train the feature extractor $\boldsymbol{\Phi}_{FE}$, and the labeled head $\boldsymbol{\Phi}_{LAB}$ with D_{lab}. When we pass an instance from D_{unlab} through $\boldsymbol{\Phi}_{LAB} \circ \boldsymbol{\Phi}_{FE}(\boldsymbol{x})$, the unlabeled instances would be predicted to one of the labeled classes consistently. We make use of these overconfident predictions from the labeled head to guide unknown identification in $\boldsymbol{\Phi}_{ULB}$. An information theoretic approach to achieve this would be to maximize the mutual information between the predictions from labeled head and unlabeled head, such that we can transfer semantic information from the labeled to unlabeled head, as motivated by [3]. Concretely, for an image $\boldsymbol{x} \in D_{unlab}$, let $\boldsymbol{l} = \boldsymbol{\Phi}_{LAB} \circ \boldsymbol{\Phi}_{FE}(\boldsymbol{x})$ denote the logits from the labeled head and $\boldsymbol{u} = \boldsymbol{\Phi}_{ULB} \circ \boldsymbol{\Phi}_{FE}(\boldsymbol{x})$ denote the logits from the unlabeled head. $\boldsymbol{l}$ and $\boldsymbol{u}$ can be of different dimensions: $\boldsymbol{l} \in \mathbb{R}^M$ and $\boldsymbol{u} \in \mathbb{R}^N$. We intend to guide the learning of $\boldsymbol{\Phi}_{ULB}$ by maximizing the mutual information $I(\boldsymbol{l}; \boldsymbol{u})$ between $\boldsymbol{l}$ and $\boldsymbol{u}$, which we can expand as follows:

$$\begin{aligned} I(\boldsymbol{l}; \boldsymbol{u}) &= H(\boldsymbol{l}) - H(\boldsymbol{l}|\boldsymbol{u}) \\ &= -\mathbb{E}_{\boldsymbol{l}}[\log p(\boldsymbol{l})] + \mathbb{E}_{\boldsymbol{l},\boldsymbol{u}}[\log p(\boldsymbol{l}|\boldsymbol{u})] \end{aligned} \tag{2}$$

where $H(\boldsymbol{l})$ refers to the entropy of $\boldsymbol{l}$ and $H(\boldsymbol{l}|\boldsymbol{u})$ is the conditional entropy between the random variables $\boldsymbol{l}$ and $\boldsymbol{u}$, sampled from a probability distribution $p(.)$. Numerically computing exact mutual information is intractable, and hence we resort to a variational approximation $q(\boldsymbol{l}|\boldsymbol{u})$ to true distribution $p(\boldsymbol{l}|\boldsymbol{u})$ [2,3] as follows:

$$\begin{aligned} I(\boldsymbol{l}; \boldsymbol{u}) &= -\mathbb{E}_{\boldsymbol{l}}[\log p(\boldsymbol{l})] + \mathbb{E}_{\boldsymbol{l},\boldsymbol{u}}[\log p(\boldsymbol{l}|\boldsymbol{u})] \\ &\approx -\mathbb{E}_{\boldsymbol{l}}[\log p(\boldsymbol{l})] + \mathbb{E}_{\boldsymbol{l},\boldsymbol{u}}[\log q(\boldsymbol{l}|\boldsymbol{u})] + \mathbb{E}_{\boldsymbol{u}}[KL(p(\boldsymbol{l}|\boldsymbol{u}) \;||\; q(\boldsymbol{l}|\boldsymbol{u}))] \\ &\geq -\mathbb{E}_{\boldsymbol{l}}[\log p(\boldsymbol{l})] + \mathbb{E}_{\boldsymbol{l},\boldsymbol{u}}[\log q(\boldsymbol{l}|\boldsymbol{u})] \end{aligned} \tag{3}$$

We assume the variational distribution to be a Gaussian, with a learnable mean function $\boldsymbol{\mu_\theta}(\boldsymbol{u})$ and variance function $\boldsymbol{\sigma_\omega}$. This would extend the derivation in Eqn. 3 to the following:

$$\begin{aligned} I(\boldsymbol{l}; \boldsymbol{u}) &\geq -\mathbb{E}_{\boldsymbol{l}}[\log p(\boldsymbol{l})] + \mathbb{E}_{\boldsymbol{l},\boldsymbol{u}}[\log q(\boldsymbol{l}|\boldsymbol{u})] \\ &= -\mathbb{E}_{\boldsymbol{l}}[\log p(\boldsymbol{l})] + \mathbb{E}_{\boldsymbol{l},\boldsymbol{u}}[\sum_{i=1}^{M} \log \sigma_\omega^i + \frac{(l^i - \mu_\theta(\boldsymbol{u}))^2}{2(\sigma_\omega^i)^2}] \end{aligned} \tag{4}$$

The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ of the mean and variance functions, the unlabeled head $\boldsymbol{\Phi}_{ULB}$ and the feature extractor $\boldsymbol{\Phi}_{FE}$ would be updated to maximize the mutual information between $\boldsymbol{l}$ and $\boldsymbol{u}$. As the first term in the RHS of Eqn. 4 is a constant, we can rewrite our mutual information based loss as below. L_{MI} is minimised along with the standard cross-entropy loss between pseudo labels and the predictions from the unlabeled head $\boldsymbol{\Phi}_{ULB}$.

$$L_{MI} = -I(\boldsymbol{l};\boldsymbol{u}) \approx -\mathbb{E}_{\boldsymbol{l},\boldsymbol{u}}[\sum_{i=1}^{M} \log \sigma_{\omega}^{i} + \frac{(l^i - \mu_{\theta}(\boldsymbol{u}))^2}{2(\sigma_{\omega}^{i})^2}] \tag{5}$$

3.4 Towards task-agnostic Inference

So far, we introduced an effective mechanism to address forgetting and an intuitive approach to enhance class discovery. Our basic architecture contains a feature extractor $\boldsymbol{\Phi}_{FE}$ which branches off into the labeled head $\boldsymbol{\Phi}_{LAB}$ and the unlabeled head $\boldsymbol{\Phi}_{ULB}$. During inference, if we know whether a sample indeed belongs to one of the labeled classes or not, we could effectively route it to the corresponding head. But, this would limit the applicability in many realistic scenarios. We circumvent this by learning a function, which we call *KCI: Known Class Identifier*, which automates this decision.

KCI is realised as a two layer neural network $\boldsymbol{\Phi}_{KCI}$ which is trained during the class discovery phase. Hence, labeled instances cannot be accessed to learn this binary function. Instead, we use the methodology explained in Sec. 3.2 to generate N_p pseudo-latents $\boldsymbol{z}_p$, which would act as a proxy for the labeled data. Using the N_u unlabeled data that we have access to, we extract their latent representations $\boldsymbol{z}_u = \boldsymbol{\Phi}_{FE}(\boldsymbol{x})$, where $\boldsymbol{x} \sim D_{unlab}$. We create a dataset of latent representations $D_{KCI} = (\mathbb{Z}_{KCI}, \mathbb{Y}_{KCI})$ where $\mathbb{Z}_{KCI} = \{\boldsymbol{z}_p^i\}_{i=1}^{N_p} \cup \{\boldsymbol{z}_u^i\}_{i=1}^{N_u}$ and $\mathbb{Y}_{KCI} = \{0\}_{i=1}^{N_p} \cup \{1\}_{i=1}^{N_u}$. We learn KCI with the following loss function:

$$L_{KCI} = \frac{1}{N_p + N_u} \sum_{i=1}^{N_p+N_u} y_i \log(\boldsymbol{\Phi}_{KCI}(\boldsymbol{z}_i)) + (1 - y_i) \log(1 - (\boldsymbol{\Phi}_{KCI}(\boldsymbol{z}_i))) \tag{6}$$

This simple formulation learns an effective classifier that differentiates labeled instances from others. We show how the learning of $\boldsymbol{\Phi}_{KCI}$ matures with training in Sec. 5.1. At inference time, given a latent representation of a test instance $\boldsymbol{z}_t$, we compute $\boldsymbol{\Phi}_{KCI}(\boldsymbol{z}_t)$ and threshold it using τ, to decide on the prediction. We include a sensitivity analysis on τ in Sec. 5.1.

4 Experiments and Results

We define the experimental protocol and evaluate our proposed methodology in this section. We formulate five different data splits across three existing datasets and benchmark against three prominent NCD approaches. We explain these in Sec. 4.1 followed by the implementation details in Sec. 4.2 and results in Sec. 4.3.

4.1 Experimental Setting

Dataset and Splits: We propose to evaluate NCDwF models on CIFAR-10 [29], CIFAR-100 [29] and ImageNet [43] datasets. Inspired by the data splits used to evaluate Novel Class Discovery methods [19,61,16], we derive a labeled set and unlabeled set from these datasets. For CIFAR-10, we group the first five classes as labeled and the rest as unlabeled. For CIFAR-100, we propose three different groupings: with the first 80, 50 and 20 classes as labeled and the rest an unlabeled. Lastly, for ImageNet, the first 882 classes are labeled and 30 classes from the remaining 118 classes (referred to as split-A in NCD methods [19,61,16]), are learned incrementally. While learning to discover novel categories, the labeled data cannot be accessed. This is an important difference when compared with the existing NCD setting. We evaluate the trained model on the test split of the corresponding datasets.
Baseline Methods: We compare our proposed approach with three recent and top performing NCD methods: RS [19], NCL [60] and UNO [16]. To ensure fair comparison with these methods, we retrain these models with code from their official repositories, adapted to our proposed incremental setting.
Evaluation Metrics: The performance on the labeled data is measured using the standard accuracy metric. Following the practice in Clustering and NCD approaches [19,16,35,49], we use clustering accuracy to measure the performance of class discovery on unlabeled data. Denoting y_i to be a prediction that the model gives for $\boldsymbol{x}_i \in D_{unlab}$, the clustering accuracy is computed as follows:

$$\text{Clustering Accuracy} = \max_{p \in perm(\mathcal{Y}_{unlab})} \frac{1}{N_u} \sum_{i=1}^{N_u} \mathbb{1}\{y_i = p(\hat{y}_i)\} \tag{7}$$

where $perm(\mathcal{Y}_{unlab})$ is a set of permutations of the unlabeled classes optimally computed via the Hungarian algorithm [30] and N_u refers to the number of instances in D_{unlab}. This discounts for the fact that the predicted cluster label might not match the exact ground truth label $\hat{y}_i$.

4.2 Implementation Details

We use ResNet-18 [21] backbone for all our experiments. We use SGD with momentum parameter of 0.9 to train the model on mini-batches of size 512. We use 200 epochs for each phase. Following our baseline [16], we also use multi-head clustering and over-clustering for the class discovery head. We strictly follow Fini *et al.* [16] for the design choice of the heads and the hyper-parameters. KCI is modeled as a two layer neural network with 128 neurons each, terminating with a single neuron. For generating the pseudo-data, we sample the mixing coefficient α from $Beta(1, 100)$. In the class discovery phase each mini-batch contains 0.25% of pseudo-data. The models are evaluated in both task-aware and task-agnostic setting. While doing task-aware inference, we assume that task identifying information (whether it belongs to any of the labeled class or not) is available with each test sample. In the more pragmatic task-agnostic

Table 2: Performance of the model in identifying instances of the labeled categories, along with identifying novel categories ('Lab' and 'Unlab' columns respectively), after incrementally learning to discover novel categories is recorded below. We note that our pseudo-data based replay and mutual information based regularization can offer improved class discovery while retaining the performance on the labeled classes, in the task-aware and generalized setting. Please find detailed description in Sec. 4.3.

Settings (→)	CIFAR-10-5-5			CIFAR-100-80-20			CIFAR-100-50-50			CIFAR-100-20-80			ImageNet-1000-882-30		
Methods (↓)	Lab	Unlab	All	Lab	Unlab	All	Lab	Unlab	All	Lab	Unlab	All	Lab	Unlab	All
Task Aware Evaluation															
RS [19]	20.00	84.48	52.24	44.1	55.7	49.9	18.14	32.56	25.35	13.05	11.5	12.28	3.34	24.54	13.94
NCL [60]	20.00	59.96	39.98	13.59	57.9	35.75	10.14	12.18	11.16	12.65	4.73	8.69	1.52	11.45	6.49
UNO [16]	33.16	**93.22**	63.19	2.01	72.78	37.39	1.76	53.85	27.81	7.95	48.7	28.33	0.75	63.4	32.08
Ours	**92.72**	90.32	**91.52**	**65.03**	**77.03**	**71.03**	**73.18**	**55.66**	**64.42**	**84.8**	**49.67**	**67.24**	**27.46**	**79.07**	**53.27**
Generalized Evaluation															
UNO [16]	0	71.36	35.68	0	58.15	29.08	0	34.22	17.11	0	41.61	20.81	0	68.34	34.17
Ours	**79.68**	**73.66**	**76.67**	**53.23**	**60.6**	**56.92**	**62.76**	**36.42**	**49.59**	**57.85**	**42.18**	**50.02**	**21.32**	**70.99**	**46.16**

setting, we use the proposed KCI to make this decision. For fair evaluation, we use KCI both with our approach and the baseline method [16]. After deciding on a specific head (either using the ground-truth or KCI), we take the *argmax* over the logits to generate the prediction. RS [19] and NCL [60] learn a binary classifier per unlabeled class, while UNO [16] and our method learn a classifier that scores via softmax function.

4.3 Results

We summarise our main results in Tab. 2. In the first row, we refer to the different data splits via the following concise notation: dataset−total_class_count−labeled_classes−unlabeled_classes. 'Lab' and 'Unlab' columns refer to the performance of the model on the labeled and the unlabeled data respectively, after learning to discover novel categories. 'All' column gives the average performance which gives a holistic measure of capacity across all classes. The first section of the table showcases the results in a task-aware setting. RS, NCL and UNO tend to forget how to detect instances from the labeled classes while trying to discover novel categories from unlabeled data. The unified head approach in UNO substantially improves the performance of class discovery. Our proposed pseudo-latent based replay mechanism, combined with MI based regularization helps to achieve improved class discovery capability while retaining the performance on the labeled classes. The forgetting is even more intense in the task-agnostic evaluation setting due to the inherent confusion caused due to absence of task identifying information. KCI helps to address this to an extent, which complements the improved performance of all the classes, when compared to the baseline.

On CIFAR-100 dataset, we experiment with changing the ratio of the labeled and unlabeled classes. We see a steady decrease in the class discovery performance and an increase in performance in recognizing instances from labeled

classes when there are lesser number of classes in the labeled pool. This implies that better pertaining on more variety of labeled classes will improve NCD.

5 Analysis

We provide additional analysis here and in supplementary materials.

5.1 Learning the Known Class Identifier

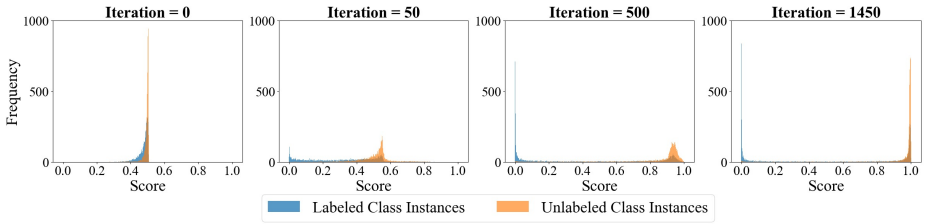

Fig. 5: KCI learns to discriminate between latent representations from unlabeled data and the pseudo latents. These plots show the classification performance on test set of labeled and unlabeled classes, showing how KCI generalizes as the learning progresses.

In Fig. 5 we visualise how the Known Class Identifier matures as the learning progresses in the CIFAR-100-50-50 setting. Before the learning starts, both the labeled and unlabeled latents are classified equally-likely. As the learning progresses, the KCI is able to disambiguate the majority of the labeled and unlabeled samples. Still, there are some false-positives which is the reason for the performance difference between task-aware and generalized evaluation in Tab. 2. We run a sensitivity analysis on τ (the threshold used to decide on the prediction from KCI) in Tab. 3. As τ increases, the performance on the labeled data increases and the unlabeled performance decreases. We use $\tau = 0.99$ throughout our experiments.

Table 3: Sensitivity analysis on the threshold τ.

Setting	CIFAR-10-5-5			CIFAR-100-80-20			CIFAR-100-50-50			CIFAR-100-20-80		
τ	Lab	Unlab	All	Lab	Unlab	All	Lab	Unlab	All	Lab	Unlab	All
0.8	64.94	80.66	72.8	47.85	70.51	59.18	47.85	51.95	49.9	42.87	48.73	45.8
0.85	66.21	79.58	72.9	48.73	69.05	58.89	48.63	51.66	50.15	44.33	48.43	46.38
0.9	69.53	77.83	73.68	49.7	69.14	59.42	50.39	50.87	50.63	45.41	47.94	46.68
0.95	72.55	74.91	73.73	52.24	66.99	59.62	53.32	49.02	51.17	49.31	46.77	48.04
0.99	79.68	73.66	76.67	53.23	60.6	56.92	62.76	36.42	49.59	57.85	42.18	50.02
0.999	85.05	51.66	68.36	61.42	45.5	53.46	69.23	27.14	48.19	65.62	34.17	49.9

5.2 On Mutual Information based Regularization

As illustrated in Fig. 6, an unlabeled instance gets misclassified into a semantically similar labeled category. This motivates us to enhance class discovery using this semantic information in the head trained on the labeled data. We couple

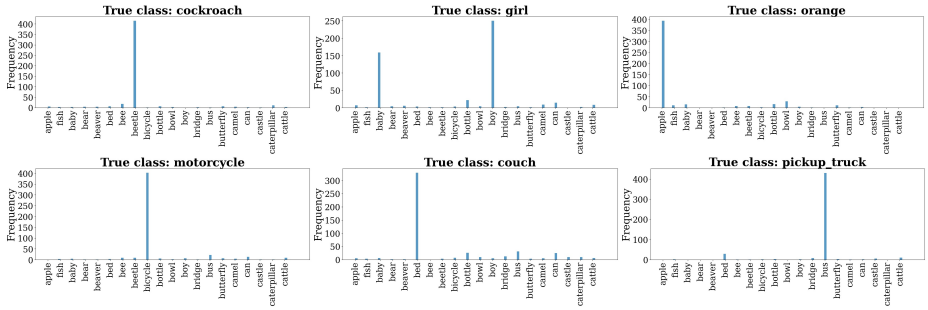

Fig. 6: The x-axis in these frequency plots represents the labeled classes in CIFAR-100-20-80 setting. Each plot shows predictions for instances of an unlabeled class (referred to as 'True class') from the labeled head. We see that most of the unlabeled instances gets misclassified into semantically meaningful labeled categories.

the mutual dependency between labeled and unlabeled heads by maximizing the mutual information between them. This helps to transfer the semantic information from the labeled to the unlabeled head, effectively guiding its class discovery capability. Such improvement is evident from the results in Tab. 2. Further, we validate the efficacy of maximizing the mutual-information between the labeled and unlabeled head in standard NCD setting too by adding it to UNO [16]. Our extra regularization is able to positively improve in this setting too, as seen in the results in Tab. 4.

Table 4: Adding Mutual Information Regularizer (MIR) to standard NCD method UNO [16].

Setting	CIFAR-10-5	CIFAR-100-80	CIFAR-100-50
UNO	94.15	89.31	59.45
UNO + MIR	**94.43**	**91.26**	**61.23**

6 Conclusion

We introduce Novel Class Discovery without Forgetting, a pragmatic extension to NCD setting. We develop an effective approach for NCDwF, which makes use of pseudo-latents as a surrogate to labeled instances to defy forgetting, and a mutual-information based regularizer to enhance class discovery. We operate in a generalized setting, where a test instance can come from any of the classes of interest. We propose to use Known Class Identifier to segregate labeled instance from the unlabeled ones during inference. We report results on five different data-splits across three datasets to test the mettle of our approach. We hope our work can shed light on this challenging problem and inspire more efforts towards this realistic setting.

Acknowledgements: KJJ was a Student Researcher at Google. We are grateful to the Department of Science and Technology, India, as well as Intel India for the financial support of this project through the IMPRINT program (IMP/2019/000250). This work is supported in part by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27208022) and HKU Startup Fund. We also thank our anonymous reviewers for their valuable feedback.

References

1. Abati, D., Tomczak, J., Blankevoort, T., Calderara, S., Cucchiara, R., Bejnordi, B.E.: Conditional channel gated networks for task-aware continual learning. In: CVPR (2020) 5
2. Agakov, D.B.F.: The im algorithm: a variational approach to information maximization. In: NeurIPS (2004) 9
3. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: CVPR (2019) 9
4. Allen, K., Shelhamer, E., Shin, H., Tenenbaum, J.: Infinite mixture prototypes for few-shot learning. In: ICML (2019) 4
5. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. In: ICLR (2020) 8
6. Belouadah, E., Popescu, A.: Il2m: Class incremental learning with dual memory. In: ICCV (2019) 5
7. Bulat, A., Kossaifi, J., Tzimiropoulos, G., Pantic, M.: Toward fast and accurate human pose estimation via soft-gated skip connections. In: IEEE International Conference on Automatic Face and Gesture Recognition (2020) 1
8. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020) 8
9. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: ECCV (2018) 4
10. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks (2009) 4
11. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. In: ICLR (2019) 4
12. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. NeurIPS (2013) 8
13. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: ICCV (2020) 4
14. Du, K.L.: Clustering: A neural network approach. Neural Networks (2010) 4
15. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: ICCV (2019) 1
16. Fini, E., Sangineto, E., Lathuilière, S., Zhong, Z., Nabi, M., Ricci, E.: A unified objective for novel class discovery. In: ICCV (2021) 3, 5, 8, 11, 12, 14
17. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences (1999) 4, 7
18. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. IEEE TPAMI (2020) 2, 4
19. Han, K., Rebuffi, S.A., Ehrhardt, S., Vedaldi, A., Zisserman, A.: Automatically discovering and learning new visual categories with ranking statistics. In: ICLR (2020) 3, 5, 11, 12
20. Han, K., Vedaldi, A., Zisserman, A.: Learning to discover novel visual categories via deep transfer clustering. In: ICCV (2019) 5
21. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016) 11
22. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015) 6

23. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: CVPR (2019) 7
24. Hsu, Y.C., Lv, Z., Kira, Z.: Learning to cluster in order to transfer across domains and tasks. In: ICLR (2018) 5
25. Hsu, Y.C., Lv, Z., Schlosser, J., Odom, P., Kira, Z.: Multi-class classification without multi-class labels. In: ICLR (2019) 5
26. Jia, X., Han, K., Zhu, Y., Green, B.: Joint representation learning and novel category discovery on single- and multi-modal data. In: ICCV (2021) 5
27. Joseph, K., Balasubramanian, V.N.: Meta-consolidation for continual learning. In: NeurIPS (2020) 4, 5
28. Joseph, K., Rajasegaran, J., Khan, S., Khan, F.S., Balasubramanian, V.N.: Incremental object detection via meta-learning. IEEE TPAMI (2021) 7
29. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto (2009) 11
30. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly (1955) 11
31. Li, Z., Hoiem, D.: Learning without forgetting. IEEE TPAMI (2017) 4
32. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS (2020) 4
33. Liu, Y., Schiele, B., Sun, Q.: Adaptive aggregation networks for class-incremental learning. In: CVPR (2021) 5
34. Liu, Y., Su, Y., Liu, A.A., Schiele, B., Sun, Q.: Mnemonics training: Multi-class incremental learning without forgetting. In: CVPR (2020) 4, 5
35. Luo, D., Ding, C., Huang, H., Li, T.: Non-negative laplacian embedding. In: ICDM (2009) 11
36. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation (1989) 4, 7
37. Mohan, R., Valada, A.: Efficientps: Efficient panoptic segmentation. IJCV (2021) 1
38. Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Processing (2014) 4
39. Rajasegaran, J., Hayat, M., Khan, S., Khan, F.S., Shao, L.: Random path selection for incremental learning. In: NeurIPS (2019) 5
40. Rajasegaran, J., Hayat, M., Khan, S., Khan, F.S., Shao, L., Yang, M.H.: An adaptive random path selection approach for incremental learning. arXiv preprint arXiv:1906.01120 (2019) 5
41. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: CVPR (2017) 4, 5
42. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: ICML (2015) 4
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015) 11
44. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016) 5
45. Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected gans converge faster. In: NeurIPS (2021) 1
46. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE TPAMI (2012) 2, 4

47. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017) 4
48. Tang, K., Miao, D., Peng, W., Wu, J., Shi, Y., Gu, Z., Tian, Z., Wang, W.: Codes: Chamfer out-of-distribution examples against overconfidence issue. In: ICCV (2021) 2
49. Tolić, D., Antulov-Fantulin, N., Kopriva, I.: A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering. Pattern Recognition (2018) 11
50. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. In: NeurIPS (2021) 1
51. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning (2020) 4
52. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. In: CVPR (2022) 5
53. Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: CVPR (2019) 4
54. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: CVPR (2017) 4
55. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: ICML (2016) 5
56. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on Neural Networks (2005) 4
57. Zhang, H., Zhan, T., Davidson, I.: A self-supervised deep learning framework for unsupervised few-shot learning and clustering. Pattern Recognition Letters (2021) 4
58. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018) 7, 8
59. Zhao, B., Han, K.: Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In: NeurIPS (2021) 5
60. Zhong, Z., Fini, E., Roy, S., Luo, Z., Ricci, E., Sebe, N.: Neighborhood contrastive learning for novel class discovery. In: CVPR (2021) 3, 5, 11, 12
61. Zhong, Z., Zhu, L., Luo, Z., Li, S., Yang, Y., Sebe, N.: Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In: CVPR (2021) 5, 11
62. Zhou, D.W., Ye, H.J., Zhan, D.C.: Learning placeholders for open-set recognition. In: CVPR (2021) 2