Spacing Loss for Discovering Novel Categories

K J Joseph[1,2] Sujoy Paul[2] Gaurav Aggarwal[2] Soma Biswas[3] Piyush Rai[2,4]
Kai Han[2,5] Vineeth N Balasubramanian[1]
[1]Indian Institute of Technology Hyderabad [2]Google Research [3]Indian Institute of Science
[4]Indian Institute of Technology Kanpur [5]The University of Hong Kong
{cs17m18p100001, vineethnb}@iith.ac.in, somabiswas@iisc.ac.in,
{sujoyp, gauravaggarwal, piyushrai, kaihanx}@google.com

Abstract

Novel Class Discovery (NCD) is a learning paradigm, where a machine learning model is tasked to semantically group instances from unlabeled data, by utilizing labeled instances from a disjoint set of classes. In this work, we first characterize existing NCD approaches into single-stage and two-stage methods based on whether they require access to labeled and unlabeled data together while discovering new classes. Next, we devise a simple yet powerful loss function that enforces separability in the latent space using cues from multi-dimensional scaling, which we refer to as Spacing Loss. Our proposed formulation can either operate as a standalone method or can be plugged into existing methods to enhance them. We validate the efficacy of Spacing Loss with thorough experimental evaluation across multiple settings on CIFAR-10 and CIFAR-100 datasets.

1. Introduction

Availability of large amount of annotated data has fueled unprecedented success of deep learning in various machine learning tasks [5, 9, 18, 30, 36, 37]. Though human learners also require various levels of supervision throughout their lifetime, we make use of the bulk of knowledge acquired so far to make intelligent choices, which guides effective learning. Drawing a parallel to the machine learning problem of image classification, it is natural to expect a model trained on a huge number of labeled classes (e.g., 1000 classes in ImageNet dataset [34]) to give meaningful representations to identify and differentiate instances of novel categories. This is the basis for the research efforts in Novel Class Discovery (NCD) setting [10, 11, 13, 15, 16, 44, 45]. Given access to labeled training data from a set of classes, an NCD model identifies novel categories from an unlabeled pool containing instances from a disjoint set of classes.

As the nascent field of Novel Class Discovery continues to evolve, we introduce a categorization of existing NCD methods based on the data that is required to train them. *Single-stage* NCD models can access labeled data and unlabeled data together while discovering novel categories from the latter. *Two-stage* NCD models can access labeled and unlabeled data only in stages. Each of these settings has a wide practical applicability. Consider a marine biologist who studies about various kinds of organisms in the ocean, from images captured by under-water vehicles [19, 20]. While analysing these images for novel categories in their lab, it would be ideal to make use of any annotated data that they might have already collected overtime. Hence, a single-stage NCD methods would be ideal for their setting. Contrastingly, consider an autonomous robot that can assist the visually impaired [24, 25]. While being operational, it would be great for the robot to discover and identify instances of novel categories in the environment, so that it can alert its users. In this scenario, it is not practical to reuse all labeled instances that the robot was trained on in its factory, while discovering novel categories. A two-stage NCD method is more desired in this setting.

A common theme in most NCD methodologies is to learn a feature extractor using the labeled data and use clustering [13, 15, 16], psuedo-labelling based learning [10, 11] or contrastive learning [17, 45] to identify classes in the unlabeled pool. In contrast, we propose a novel *Spacing Loss* which ensures separability in the latent space of feature extractor, for the labeled and unlabeled classes. This is achieved by transporting semantically dissimilar instances to equidistant areas in the latent space, identified via multi-dimensional scaling [40]. We note that our proposed loss formulation is orthogonal to the existing methodologies, and can easily complement these methods. Our experimental evaluation on CIFAR-10 [23] and CIFAR-100 [23] datasets suggests that the models trained with the proposed Spacing Loss achieve state-of-the-art performance when compared to two-stage NCD methods. Further, when combined with single-stage methodologies, our loss formulation improves each of them consistently.

The standard strategy to evaluate NCD methods is to train the model on a subset of classes from a classification dataset and evaluate its performance on the remaining classes. Complementing existing protocols, we introduce a new split where the number of classes in the labeled pool is significantly lower than the number of classes in the unlabeled pool. Such a protocol aligns more closely with the real-world scenarios, where the number of classes in the labeled and unlabeled pool might be heavily imbalanced.

To summarize, the key contributions of our work are:

- We propose Spacing Loss, which enforces separability in the latent space, for the challenging problem of novel category discovery.
- We evaluate our proposed approach on benchmark datasets for novel category discovery, under both single- and two-stage settings, consistently outperforming existing methods.

2. Novel Class Discovery Methods

Two-stage Methods Early methods in Novel Class Discovery [13, 15, 16] operate in a phased setting. In the first phase, the model learns from the labeled data, and in the subsequent phase, it discover novel categories from the unlabeled pool. MCL [16] and KCL [15] learn a binary similarity function using meta-learning in the first phase, and use this in the category discovery phase. DTC [13] first learns a feature extractor on the labeled data. In the next stage, these features are used to initialise a clustering algorithm [42], which further fine-tunes these representations using the unlabeled data, thereby improving class discovery.

Single-stage Methods More recent efforts in NCD [10, 11, 45] use labeled and the unlabeled data together to discover novel categories. RS [11, 12], NCL [45] and OpenMix [46] first use RotNet [22] to self-supervise on the labeled and unlabeled data. Then, RS [11] uses pseudo-labels from ranking-statistics method to learn an unlabeled head. NCL [45] and Jia *et al.* [17] find that contrastive learning improves class discovery and OpenMix [46] uses mix-up [43] to generate more training data to guide class discovery. UNO [10] finds that a unified loss function enhances the synergy between the learnings from labeled and the unlabeled data. Zhao and Han [44] proposes to focus on fine-grained local cues in images to enhance discrimination[1].

3. Spacing Loss

Learning to adapt the latent representations of a model, such that semantically identical samples would share nearby locations in the latent manifold, while semantically dissimilar samples are spaced apart, would be ideal for discovering novel classes. Such a subspace shaping should evolve as latent representations mature. Two characteristics would be ideal in such a setting: 1) the ability to transport similar samples to locations equidistant from other dissimilar samples in the latent manifold, 2) the datapoints having the ability to refresh their associativity to a group as the learning progresses. We propose a simple yet effective methodology

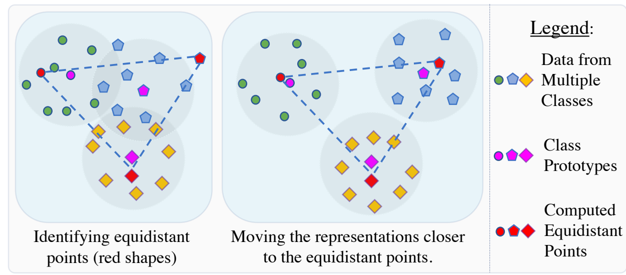

Figure 1. The figure illustrates how latent space is adapted by the proposed Spacing Loss. Latent representations from different classes are shown in different shapes. As the model is bootstrapped with labeled data, the latent representations from the unlabeled data will have reasonable semantic grouping. We further enhance the separability in the latent space by identifying equidistant points (shown in red) and then moving the latent representations to these identified locations, effectively ensuring spacing between the classes of interest.

that accommodates both aspects. Fig. 1 illustrates how the latent space is adapted using the proposed Spacing Loss. While learning to discover classes, we identify locations in latent space (in red), which are equidistant from each other. Next, we enforce the latent representations from unlabeled data to be transported to nearest of such points. Each latent representation can change their membership to a specific group as the learning progresses. This flexibility along with the weak regularization enables us to learn a well-separated latent representation. A simple non-parametric inference in this space can help us to discover categories. We summarize how equidistant locations is identified in Sec. 3.1, followed by how latent space is adapted in Sec. 3.2, concluding with the overall objective in Sec. 3.3.

3.1. Finding Equidistant Points in the Latent Space

Let us consider a feature extractor $\Phi : \mathbb{R}^{w \times h \times 3} \rightarrow \mathbb{R}^z$, which takes an input image and generates a z dimensional latent representation. We identify c prototypes, $\boldsymbol{P} = \{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_c\}$, from these latent representations, where c is the total number of classes under consideration. These prototypes can be initialised using a simple centroid based strategy. We identify equidistant points, $\boldsymbol{P}^e = \{\boldsymbol{p}_1^e \cdots \boldsymbol{p}_c^e\}$, in this latent space which are guaranteed to be far apart atleast by the largest pair-wise distance between these prototypes. These equidistant points serve as anchors to which the corresponding centroids and its associated data would be progressively shifted to while the learning progresses.

Let p_{dist} be the largest pair-wise distance between the prototypes. We first construct a $c \times c$ *dissimilarity matrix* $\boldsymbol{\Delta}$ as follows: all entries but for the diagonals are set to $\delta_{ij} = \alpha \times p_{dist}$, where $\alpha > 1$. The diagonal elements δ_{ij}, of $\boldsymbol{\Delta}$ are set to 0. Hence, $\boldsymbol{\Delta}$ is symmetric, non-negative and hollow by construction. Given a dataset $\boldsymbol{D}$, we seek to find $\boldsymbol{P}^e = \{\boldsymbol{p}_1^e \cdots \boldsymbol{p}_c^e\}$ where each $\boldsymbol{p}_i^e \in \mathbb{R}^z$, such that distance between any $\boldsymbol{p}_i^e$ and $\boldsymbol{p}_j^e$ is approximately δ: ie. $d_{ij}(\boldsymbol{P}^e) \approx$

[1] As NCD is a nascent field, we will maintain an updated list of methods here: https://github.com/JosephKJ/Awesome-Novel-Class-Discovery.

δ_{ij}. $d_{ij}(\boldsymbol{P}^e)$ corresponds to the distance between $\boldsymbol{p}_i^e$ and $\boldsymbol{p}_j^e$ in euclidean space. We can formulate the objective to learn $\boldsymbol{P}^e$ as follows:

$$\sigma(\boldsymbol{P}^e) = \sum_{i<j\leq c} w_{ij}(d_{ij}(\boldsymbol{P}^e) - \delta_{ij})^2, \tag{1}$$

where $\boldsymbol{W}$ is a symmetric, non-negative and hollow matrix of weights w_{ij}, which captures the relative importance. For simplicity, we weigh each $\boldsymbol{P}_i^e$ equally. As finding an analytical solution to minimize Eq. (1) is intractable, an iterative majorization algorithm [4, 7, 40] is used. We seek to find a manageable surrogate function $\tau(\boldsymbol{P}^e, \boldsymbol{Y})$, which majorizes $\sigma(\boldsymbol{P}^e)$, i.e., $\tau(\boldsymbol{P}^e, \boldsymbol{Y}) > \sigma(\boldsymbol{P}^e)$, with the initial supporting points $\boldsymbol{Y}$. We can rewrite Eq. (1) as follows:

$$\sigma(\boldsymbol{P}^e) = \sum_{i<j} d_{ij}^2(\boldsymbol{P}^e) + \sum_{i<j} \delta_{ij}^2 - 2\sum_{i<j} \delta_{ij} d_{ij}(\boldsymbol{P}^e). \tag{2}$$

The first term is a quadratic in $\boldsymbol{P}^e$ and can be expressed as $\mathrm{Tr}\, \boldsymbol{P}^{eT} \boldsymbol{V} \boldsymbol{P}^e$, where $\boldsymbol{V}$ has $v_{ij} = -w_{ij}$ and $v_{ii} = \sum w_{ij}$ [7]. The second term is a constant, say k, and the third term can be bounded as follows:

$$\begin{aligned} \sum_{i<j} \delta_{ij} d_{ij}(\boldsymbol{P}^e) &= \mathrm{Tr}\, \boldsymbol{P}^{eT} \boldsymbol{B}(\boldsymbol{P}^e) \boldsymbol{P}^e \\ &\geq \mathrm{Tr}\, \boldsymbol{P}^{eT} \boldsymbol{B}(\boldsymbol{Y}) \boldsymbol{Y}, \end{aligned} \tag{3}$$

where $\boldsymbol{B}(\boldsymbol{Y})$ has

$$b_{ij} = \begin{cases} \frac{\delta_{ij}}{d_{ij}(\boldsymbol{Y})}, & \text{for } d_{ij}(\boldsymbol{Y}) \neq 0, i \neq j \\ 0, & \text{for } d_{ij}(\boldsymbol{Y}) = 0, i \neq j \end{cases} \quad \text{and}$$

$$b_{ii} = -\sum_{j=1, j\neq i}^{c} b_{ij}. \tag{4}$$

The proof of this inequality follows [4, 7]. Hence, the surrogate function that majorizes $\sigma(\boldsymbol{P}^e)$ is as follows:

$$\tau(\boldsymbol{P}^e, \boldsymbol{Y}) = \mathrm{Tr}\, \boldsymbol{P}^{eT} \boldsymbol{V} \boldsymbol{P}^e + k - 2\ \mathrm{Tr}\, \boldsymbol{P}^{eT} \boldsymbol{B}(\boldsymbol{Y}) \boldsymbol{Y}. \tag{5}$$

Algorithm 1 GETEQUIDISTANTPOINTS

Input: Prototype vectors: $\boldsymbol{P} = \{\boldsymbol{p}_0 \cdots \boldsymbol{p}_c\}$, Small constant ϵ.
Output: Equidistant points: $\boldsymbol{P}^e$
1: $p_{dist} \leftarrow$ maximum distance between all prototypes in $\boldsymbol{P}$.
2: Compute $\boldsymbol{\Delta}$ from p_{dist}.
3: Initialize $\boldsymbol{P}^e$ randomly.
4: **do**
5: $\quad \boldsymbol{Y} \leftarrow \boldsymbol{P}^e$
6: $\quad \boldsymbol{P}^e \leftarrow \arg\min_{\boldsymbol{P}^e} \tau(\boldsymbol{P}^e, \boldsymbol{Y})$ ▷ Defined in Eq. (5)
7: **while** $(\boldsymbol{Y} - \boldsymbol{P}^e) > \epsilon$
8: **return** $\boldsymbol{P}^e$

Algorithm 1 summarizes how $\boldsymbol{P}^e$ are computed by optimizing Eq. (5). In Line 2, we compute the dissimilarity matrix $\boldsymbol{\Delta}$ by using the maximum distance between the prototype vectors $\boldsymbol{P}$. First $\boldsymbol{P}^e$ is randomly initialised. Until there is negligible change ϵ in $\boldsymbol{P}^e$, we update $\boldsymbol{P}^e$ to optimize the surrogate function $\tau(\boldsymbol{P}^e, \boldsymbol{Y})$. The resulting vectors in $\boldsymbol{P}^e$ are guaranteed to be equidistant from each other [4].

3.2. Learning Separable Latent Space

Once the equidistant locations in the latent space $\boldsymbol{P}^e$ are identified, they can be used to enforce separation in the latent representations of images from different classes. As each latent representation matures with training, it might need to change its associativity with its initial group. We propose a novel formulation in Algorithm 2 that would allow for this flexibility during learning. The training essentially alternates between learning with pseudo-labels derived from class prototypes (Lines 6 - 8) and modifying the class prototypes themselves (Lines 11 - 15). In Line 1, we initialize the class prototypes $\boldsymbol{P}$ as the centroids of latents from $\Phi_{\boldsymbol{\theta}}$ using k-means [29]. Based on the closeness to these prototypes, the class associativity of each image in a mini-batch is determined in Line 7. The feature extractor is updated to make the latent representations closer to these prototypes in Line 8. Using these newer features, the assignment is recomputed and the prototypes themselves are updated in Line 15. For each data-point z_i, its corresponding prototype $\boldsymbol{p}_{c_{\boldsymbol{z}_i}}$ is moved closer to the equidistant point $\boldsymbol{p}_{c_{\boldsymbol{z}_i}}^e$ and its current representation, controlled by a momentum parameter η. The parameter η dampens with more instances of the specific class seen during training.

Algorithm 2 LEARNINGWITHSPACING

Input: Feature extractor: $\Phi_{\boldsymbol{\theta}}$, Data: $\boldsymbol{D} = \{\boldsymbol{X}_i\}$, # of epochs: e.
1: Initialize class prototypes $\boldsymbol{P} = \{\boldsymbol{p}_0 \cdots \boldsymbol{p}_c\}$.
2: Identify equidistant points $\boldsymbol{P}^e = \{\boldsymbol{p}_0^e \cdots \boldsymbol{p}_c^e\}$ using Algo. 1.
3: Initialize assignment frequency $\boldsymbol{v} \leftarrow \boldsymbol{0}$; $|\boldsymbol{v}| = c$.
4: **for** each epoch e **do**
5: $\quad$ **for** each minibatch $\boldsymbol{X} \subset \boldsymbol{D}$ **do**
6: $\quad\quad \boldsymbol{Z} \leftarrow \Phi_{\boldsymbol{\theta}}(\boldsymbol{X})$
7: $\quad\quad \boldsymbol{A} \leftarrow$ assign the nearest prototype from $\boldsymbol{P}$ for each $\boldsymbol{Z}$.
8: $\quad\quad$ Update $\boldsymbol{\theta}$ with MeanSquaredError($\boldsymbol{Z}$, $\boldsymbol{A}$).
9: $\quad\quad \boldsymbol{Z} \leftarrow \Phi_{\boldsymbol{\theta}}(\boldsymbol{X})$ ▷ Recompute $\boldsymbol{Z}$ with updated $\boldsymbol{\theta}$
10: $\quad\quad \boldsymbol{A} \leftarrow$ recompute prototype assign. for each new $\boldsymbol{Z}$.
11: $\quad\quad$ **for** $\boldsymbol{z}_i$ in $\boldsymbol{Z}$ **do**
12: $\quad\quad\quad c_{\boldsymbol{z}_i} \leftarrow$ retrieve assignment index of $\boldsymbol{z}_i$ from $\boldsymbol{A}$.
13: $\quad\quad\quad \boldsymbol{v}[c_{\boldsymbol{z}_i}] \leftarrow \boldsymbol{v}[c_{\boldsymbol{z}_i}] + 1$
14: $\quad\quad\quad \eta \leftarrow \frac{1}{\boldsymbol{v}[c_{\boldsymbol{z}_i}]}$
15: $\quad\quad\quad \boldsymbol{p}_{c_{\boldsymbol{z}_i}} \leftarrow (1-\eta)\boldsymbol{p}_{c_{\boldsymbol{z}_i}} + \eta(\boldsymbol{z}_i + \boldsymbol{p}_{c_{\boldsymbol{z}_i}}^e)$

3.3. Overall Objective

So far, we have explained how the feature extractor $\Phi_{\boldsymbol{\theta}}$ is adapted by Spacing Loss. Our complete model extends this backbone with one head for the labeled data $F_{Lab} = \Phi_{Lab} \circ \Phi_{\boldsymbol{\theta}}$ and another for the unlabeled data $F_{Ulab} = \Phi_{Ulab} \circ \Phi_{\boldsymbol{\theta}}$. F_{Lab} is learned with the labeled examples. F_{Ulab} is learned with pairwise pseudo labels derived from cosine-similarity [45] between its latent repre-

Setting→	Imbalanced Class Split				Balanced Class Split			
Dataset Splits→	CIFAR-100-80-20		CIFAR-100-20-80		CIFAR-10-5-5		CIFAR-100-50-50	
Method	CA	NMI	CA	NMI	CA	NMI	CA	NMI
RS [11]	69.39	0.6934	16.63	0.4493	89.72	0.7724	47.72	0.5666
RS + Spacing loss	**73.16**	**0.7252**	**26.37**	**0.4562**	**89.90**	**0.7764**	**48.20**	**0.5712**
NCL [45]	81.01	0.7883	19.82	0.4570	92.70	0.8233	56.71	0.6355
NCL + Spacing loss	**85.11**	**0.7896**	**35.60**	**0.5064**	**93.32**	**0.8364**	**57.36**	**0.6432**

Table 1. We study the class discovery performance of single-stage NCD models across multiple settings in this table. Our proposed loss formulation can act as an add-on to existing methods, effectively enhancing their class discovery capability, even for severely skewed class distributions.

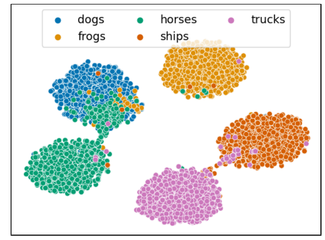

Figure 2. Latent space of novel categories from CIFAR-10-5-5, trained using NCL + Spacing Loss.

sentations. We also enforce consistency in prediction with an augmented view of each image [11,13,44,45] to enhance learning. While learning a two-stage model, we first learn F_{Lab} using cross entropy loss with labeled data and then learn F_{Ulab} with these auxiliary losses and Spacing Loss operating in the latent space. Labeled and unlabeled data, along with all the losses are used to learn the single-stage model. During inference, we do a k-means [29] on the latent representations from the backbone network, to discover novel categories.

4. Experiments and Results

Following existing NCD methods [10, 11, 13, 15, 16, 44, 45], we define splits on CIFAR-10 and CIFAR-100 to evaluate the efficacy of our method. Clustering Accuracy [11] and NMI [39] are used as the evaluation criteria. We use ResNet-18 [14] backbone and closely follow the hyper-parameter settings from Zhong *et al*. [45].

4.1. Two-stage Results

In the first phase, we train the model on the labeled data from the first 5 and 80 classes from CIFAR-10 and CIFAR-100 datasets respectively for 200 epochs. In the next phase, classes are identified from the unlabeled data guided by the Spacing Loss. Tab. 2 showcases the results. Our method consistently outperforms existing methods by a large margin, showing the efficacy of the proposed Spacing Loss. RS [11] and NCL [45] are adapted to the two-stage setting for fair comparison (denoted by *).

Datasets →	CIFAR-10		CIFAR-100	
Method	CA	NMI	CA	NMI
K-means [29]	65.5	0.422	66.2	0.555
KCL [15]	66.5	0.438	27.4	0.151
MCL [16]	64.2	0.398	32.7	0.202
DTC [13]	87.5	0.735	72.8	0.634
RS* [11]	84.6	0.658	69.5	0.581
NCL* [45]	60.5	0.479	59.5	0.428
Spacing Loss	**90.5**	**0.787**	**80.62**	**0.719**

Table 2. Regularization induced by Spacing Loss has better class discovery ability compared to baseline two-stage methods.

4.2. Single-stage Results

A key characteristic of our proposed Spacing Loss is that the latent space regularization that it offers can effectively act as an add-on to existing methodologies. We showcase this capability while evaluating in single-stage setting. In Tab. 1, we organise different dataset splits based on the balance between the number of classes in labeled and unlabeled pool. The concise notation in Row 2 can be expanded as: dataset−total_class_count−labeled_classes−unlabeled_classes. The latent space separation induced by Spacing Loss helps to improve the class discovery capability on all settings. It is interesting to note that the improvement is more pronounced in the more pragmatic setting, where the split of classes between the labeled and unlabeled pool is skewed. t-SNE [38] visualization of backbone features in Fig. 2 shows good separation in these latent representations of novel categories in CIFAR-10-5-5 setting.

5. Enhancing Continual Learning with NCD

Continual learning setting aims to learn a single model which can incrementally accumulate knowledge across multiple tasks, without forgetting. Main-stream efforts in Continual Leaning [1,2,6,8,21,26–28,28,31–33,33,35,41] assume that the data which is introduced in each incremental task is fully annotated. Efforts in Novel Class Discovery can help to relax this requirement, where the model could be tasked to identify classes from the instances of a new task automatically, based on the learnings that it already had. Then, these identified novel categories may be incrementally learned. We hope that the unification of these two streams of research would lead to a more pragmatic problem setting by building on their complementary characteristics.

6. Conclusion

We characterise research efforts in the nascent Novel Class Discovery setting into single-stage and two-stage methods, based on their data requirement during training. We further propose a simple yet effective method which enhances both these settings by enforcing separability in the latent representations. Our experimental analysis on multiple settings on two benchmark datasets corroborates with our assertions. Advancements in NCD can help continual learning models to operate in an open-world [3,18], where it can automatically identify novel categories and then incrementally learn them. We hope this pragmatic setting would be extensively explored in follow-up works.

References

[1] Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *CVPR*, pages 3931–3940, 2020. 4

[2] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *ICCV*, pages 583–592, 2019. 4

[3] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, pages 1893–1902, 2015. 4

[4] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005. 3

[5] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. In *IEEE FG*, pages 8–15, 2020. 1

[6] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, pages 233–248, 2018. 4

[7] Jan De Leeuw. Applications of convex analysis to multidimensional scaling. *Department of Statistics, UCLA*, 2005. 3

[8] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ICCV*, pages 86–102, 2020. 4

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019. 1

[10] 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, 4

[11] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *ICLR*, 2020. 1, 2, 4

[12] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE TPAMI*, 2021. 2

[13] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, pages 8401–8409, 2019. 1, 2, 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645, 2016. 4

[15] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *ICLR*, 2018. 1, 2, 4

[16] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019. 1, 2, 4

[17] 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

[18] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, pages 5830–5840, 2021. 1, 4

[19] Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Benjamin Woodward, and Katy Croff Bell. Fathomnet: A global underwater image training set for enabling artificial intelligence in the ocean. *arXiv:2109.14646*, 2021. 1

[20] Kakani Katija, Paul LD Roberts, Joost Daniels, Alexandra Lapides, Kevin Barnard, Mike Risi, Ben Y Ranaan, Benjamin G Woodward, and Jonathan Takahashi. Visual tracking of deepwater animals using machine learning-controlled robotic underwater vehicles. In *WACV*, pages 860–869, 2021. 1

[21] Joseph KJ and Vineeth Nallure Balasubramanian. Metaconsolidation for continual learning. *NeurIPS*, 2020. 4

[22] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2

[23] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Citeseer*, 2009. 1

[24] Vladimir Kulyukin, Chaitanya Gharpure, and John Nicholson. Robocart: Toward robot-assisted navigation of grocery stores by the visually impaired. In *IEEE ROS*, pages 2845–2850, 2005. 1

[25] Vladimir Kulyukin, Chaitanya Gharpure, John Nicholson, and Grayson Osborne. Robot-assisted wayfinding for the visually impaired in structured indoor environments. *Autonomous robots*, 2006. 1

[26] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 2017. 4

[27] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2544–2553, 2021. 4

[28] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *ICCV*, pages 12245–12254, 2020. 4

[29] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967. 3, 4

[30] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *IJCV*, 2021. 1

[31] Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for incremental learning. *NeurIPS*, 2019. 4

[32] Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Ming-Hsuan Yang. An adaptive random path selection approach for incremental learning. *arXiv:1906.01120*, 2019. 4

[33] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 4

[34] 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. *IJCV*, 2015. 1

[35] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv:1606.04671*, 2016. 4

[36] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *NeurIPS*, 2021. 1

[37] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 2021. 1

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 4

[39] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR*, 2010. 4

[40] Andrew R Webb. Multidimensional scaling by iterative majorization using radial basis functions. *PR*, 1995. 1, 3

[41] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, pages 374–382, 2019. 4

[42] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, pages 478–487, 2016. 2

[43] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[44] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *NeurIPS*, 2021. 1, 2, 4

[45] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *CVPR*, pages 10867–10875, 2021. 1, 2, 3, 4

[46] 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*, pages 9462–9470, 2021. 2