Self-Supervised Features Improve Open-World Learning

Akshay Raj Dhamija Touqeer Ahmad Jonathan Schwan
Mohsen Jafarzadeh Chunchun Li Terrance E. Boult
Vision and Security Technology Lab, University of Colorado at Colorado Springs, Colorado Springs
{adhamija, touqeer, jschwan2, mjafarzadeh, cli, tboult}@vast.uccs.edu

Abstract

This is a position paper that addresses the problem of Open-World learning while proposing for the underlying feature representation to be learnt using self-supervision. We also present an unifying open-world framework combining three individual research dimensions which have been explored independently i.e. *Incremental Learning, Out-of-Distribution detection and Open-World learning. We observe that the supervised feature representations are limited and degenerate for the Open-World setting and unsupervised feature representation is native to each of these three problem domains. Under an unsupervised feature representation, we categorize the problem of detecting unknowns as either Out-of-Label-space or Out-of-Distribution detection, depending on the data used during system training versus system testing. The incremental learning component of our pipeline is a zero-exemplar online model which performs comparatively against state-of-the-art on ImageNet-100 protocol and does not require any back-propagation or retraining of the underlying deep-network. It further outperforms the current state-of-the-art by simply using the same number of exemplars as its counterparts. To evaluate our approach for Open-World learning, we propose a new comprehensive protocol and evaluate its performance in both Out-of-Label and Out-of-Distribution settings for each incremental stage. We also demonstrate the adaptability of our approach by showing how it can work as a plug-in with any of the recently proposed self-supervised feature representation methods.*

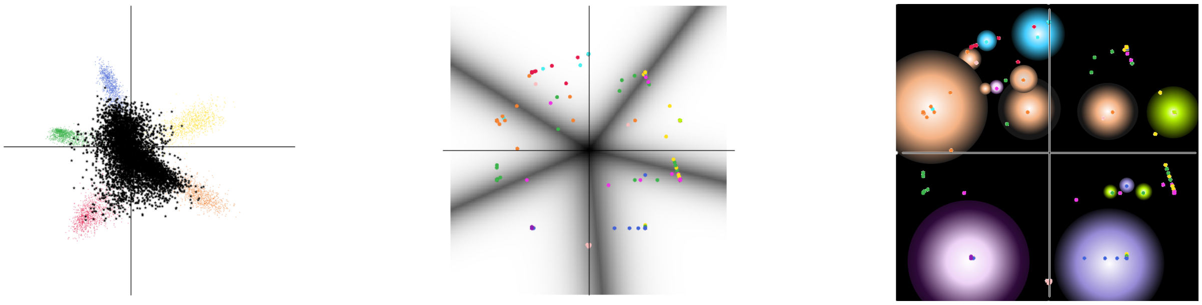
1. Introduction

Vision systems rarely operate in a closed-world where only the objects seen during training (known objects) are presented to them. Open-world learning, formalized in [6], consists of classifying the known classes, detecting unknown classes, and incrementally learning new classes. However, that paper and subsequent papers on the topic [54], have ignored the importance of feature representations, which has been studied in broader settings [31, 44, 19, 33]. In parallel, there has been a growing amount of literature on incremental or life-long learning that seeks to evolve feature representation in parallel with new classes [37, 52, 58, 30, 39, 17, 29]. The detection of unknown classes has become of growing interest, and is sometimes called Out-Of-Distribution (OOD) detection [28, 37, 56, 53, 36, 50, 7, 38, 18]. It is natural to try to combine these three research directions, to produce a true open-world visual learning system. Fig.1(a) highlights the problem with open-world learning by evolving the supervised feature-space from a few known classes. With supervised training for known classes the resulting feature space over-specializes as noticed by others as well [22]. This over-specialization limits the ability to detect out-of-distribution (unknown) classes because of their projection on top of known classes, see [16]. This weak feature representation leads to unknowns being classified as knowns. Since, the unknowns can not be identified correctly, the system would not be attempting to get labels for them and hence failing at its open-world learning task.

We contend that before operating in an open-world, an agent may gather large amounts of unlabeled data, without human intervention which may then be used to learn a feature space using self-supervised representation learning. This is analogous to how children operate in the world before ever needing to learn "semantic classes." Their representation of the world is largely learnt with self-supervision. With labels being provided for some of the objects encountered in the open-world, the system can then learn to identify them without confusing with other objects for which it had never received the labels, due to its robust self-supervised feature representation. This converts the problem of out-of-distribution detection to a much simpler out-of-label-space detection – making open-world learning much more effective, since the unknowns to be identified are within the distribution of the trained feature-space.

We argue that self-supervised open-world recognition is the natural formulation of the open-world problem and that the need to incrementally learn feature spaces as well as

https://github.com/Vastlab/SSFiOWL



(a) Representation overlap in supervised feature space

(b) SoftMax decision boundaries for known classes in the self-supervised feature space.

(c) Limiting Open-Space risk in self-supervised feature space

Fig. 1: WHY WE RECOMMEND SELF-SUPERVISED FEATURES FOR OPEN-WORLD LEARNING? *(a) shows supervised feature space while (b), (c) show probabilities from a self-supervised space (learned by [32]) for the MNIST data. In (a) we show a supervised LeNet++ trained for 5 known classes (0,1,2,3,4) in various colors. Unknowns (classes 6-9) are overlaid in black showing the problem of Out-Of-Distribution (OOD) classes overlapping and being difficult to reject as is needed for Open World learning. Our approach combines self-supervised feature learning with techniques for detecting Out-Of-Label-space (OOL). The self-supervised feature space has been learnt without labels, hence when known classes (0-4) are learnt, (6-9) are still out of label space, but mapped better. While a supervised learning like softmax may still be able to classify knowns using the decision boundaries in (b) it will still classify unknowns as knowns with a high confidence since they lie in the white space (high knowness score). Our approach (c) uses distance based metrics to identify unknown samples as it by default considers the region not around any of the known samples as unknown, hence providing a larger black are where unknowns can lie. Details in Sec 3.*

classifiers from a small set of samples is a different and more specialized *"new-world" recognition problem*. New-world recognition problems do occur naturally, e.g., when imaging modalities change such as microscopes to telescopes, MRIs to CT scans or when we enter a radically new environment such as under-water. Even more subtle changes in imaging, e.g., object-centric such as ImageNet [55] style images, vs scene-centric imaging found in the Places dataset[62], can yield sufficiently different "world views" so as to benefit from different feature spaces. But even in a new-world recognition problem, a pretrained feature space plus domain-adaption techniques should be the default approach, and supervised training of feature-spaces from only a small set of classes is mostly of academic interest to understand the feature learning process.

Our Contributions ***(a)*** A new approach for Non-backpropagating Open World Learning (Sec 3). ***(b)*** Demonstrating that self-supervised features provide superior incremental class learning even when compared to SOTA supervised learning techniques. (Tab 1) ***(c)*** Our approach that can be improved using any future advances in self-supervised learning. (Tab 2) ***(d)*** An open-world learning protocol (Sec 4) ***(e)*** Demonstrating that techniques for out-of-label space detection on self-supervised features are superior to out-of-distribution detection on supervised feature spaces.(Tab 3)

2. Background & Related Work

The goal of an open-world learner is to not only identify unknown samples but also to learn to classify the classes they belonged to in future encounters. While these tasks have been independently researched as the domain of Out-of-Distribution (OOD) detection and under Incremental Learning (IL), rarely attempts have been made to combine the two in a unified Open-World (OW) framework. In this section, we review some of the methods from the *Incremental Learning* and *Out-of-Distribution Detection* domains while also briefly discussing the related work to Open-World (OW) learning, some frameworks and their shortcomings.

2.1. Incremental Learning

In an Open-World setup, not all classes are available at the initial training phase of a learner; new classes and new instances of old classes are encountered in a temporal manner. An open-world learner should incrementally learn the new classes while maintaining its performance in classifying the old classes. Earlier approaches towards incremental learning can be grouped into the following two main categories.

Fixed Feature Representation Incremental learning approaches belonging to this category aim at classifying all the available classes without changing the representations. Nearest Mean Classifier (NMC) [43, 42] is the principal representative of incremental methods that represent each

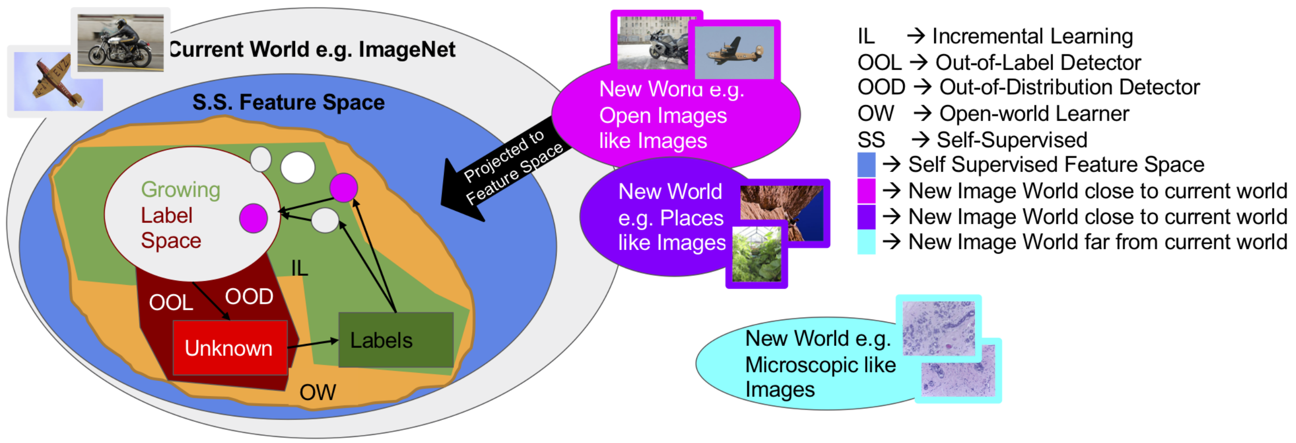

Current World e.g. ImageNet
S.S. Feature Space
Growing Label Space
IL
OOL
OOD
Unknown
Labels
OW
Projected to Feature Space
New World e.g. Open Images like Images
New World e.g. Places like Images
New World e.g. Microscopic like Images
IL → Incremental Learning
OOL → Out-of-Label Detector
OOD → Out-of-Distribution Detector
OW → Open-world Learner
SS → Self-Supervised
→ Self Supervised Feature Space
→ New Image World close to current world
→ New Image World close to current world
→ New Image World far from current world

Fig. 2: UNDERSTANDING LABEL SPACES AND CURRENT RESEARCH *Challenging the existing research directions, we argue that for Open-World learning, the native feature space to use should be self-supervised which will improve Incremental Learning as well as transform the novelty detection problem from the classic but difficult Out-of-Distribution detection to a more natural and easier Out-of-Label-Space detection. The operating-world of an agent (gray circle) is naturally sampled, from which an agent learns a feature space (blue oval) in a self-supervised manner. At the point of being applied for inference in the open world, the agent can start to learn its label-space or operating-world with a few labeled examples (batch 0 in the conventional incremental learning terminology). The label-space (white oval) can either be initialized with data from the world on which the self-supervised feature space is learned or data from another world projected to the learned feature space. The agent then incrementally discovers unknowns, gets labels for them from a teacher/annotator, and learns them, incrementally growing its label-space. Known/unknown in open-world learning is about labels used for classifier learning, not about the data used for representation learning. We differentiate Out-of-Label-Space (OOL) and Out-of-Distribution(OOD), based on the operating world vs the world from which training data was sampled. For example, if the label-space (big white oval) uses features trained with unlabeled data from the native-world (gray circle), such unknowns would be Out-of-Label space. But if the agent is initialized with features using data from a new-world (pink circle), then unknowns would all be Out-of-Distribution. It should be noted that the underlying agent is incapable of discriminating between OOL and OOD; it only detects novel samples as unknown. When the label-space is initialized with the native-world, the chance of new concepts (Out-of-Label-Space samples) overlapping other objects within the existing label-space is low, whereas it is higher when the label-space is initialized with samples from the new world.*

class using a prototype vector that is the mean of all the examples seen for that class. In DeeSIL [2], Belouadah & Popescu proposed a deep-shallow merging where fixed feature representation is used as a feature extractor, and shallow independent classifiers (SVMs) are learned to increase the recognition capacity of the model. Approaches such as DeepSLDA [25] also belong to this category. While these approaches are faster than their counterpart explained below, their performance is limited by the performance of the underlying feature representation space. As discussed in Fig 1, the performance of underlying feature-space when trained in a supervised manner has major disadvantages. While our incremental learning approach (NIL) still belongs to this small and dying subset of incremental learning approaches, we outperform the current supervised approaches by leveraging advances in self-supervised learning.

Altering Feature Representations Most recent attempts towards incremental learning try to address ***catastrophic forgetting*** and ***concept drift*** by partially/fully re-training the network. In Learning without Forgetting (LwF) [37], Li and Hoiem tried to address catastrophic forgetting by introducing knowledge distillation term in the loss function by which they encouraged network output for new classes to be close to the original network. Another common approach is maintaining *exemplars i.e*., number of samples/classes from old classes on which the network in the previous iteration was trained. The network for the next iteration is then trained not only on the new classes but also refined with the exemplars *e.g*., [52, 58, 26]. Generally, the memory budget or a number of examples/class allowed to be maintained as exemplar is fixed. Choosing these exemplars is also an active research area with methods like *herding* [57] and *mnemonics* [39] being proposed. PODNet [17] studied incremental learning as *rigid-plasticity trade-off* where the network learns to balance between remembering the old classes (*rigidity*) and learning new ones (*plasticity*). Unlike other methods [39, 58, 30] which typically employ iCaRL protocol [52] where 5,10 or more classes are introduced per incremental-task, in [17], PODNet and existing approaches [52, 29, 58] are evaluated on learning one class per task. Authors demonstrated that PODNet fights catastrophic forgetting better than the earlier approaches over these very long runs of small incremental tasks. By introducing the accommodation ratio to the cross-distillation loss, He *et al*. [26] tried to address the incremental learning in a more challenging online setting where update time and available data are limited. To achieve lifelong learning, each online incremental learning phase is followed by an offline retraining phase where all the data available up to that point is used to retrain the network. Like others, they also maintain an exemplar set and employ herding [57] to choose the examples for each class. Recently, there have been more attempts towards incremental learning [61, 51] as well as surveying and reviewing the literature on the topic [40, 41]. In [5], Belouadah *et al*. conducted a comprehensive evaluation of recently proposed incremental learning approaches [37, 52, 29, 58, 3, 4, 2, 8, 24, 25] and concluded that none of the existing algorithms was better than the others; both memory and incremental step size influence the relative performance.

2.2. Out-of-Distribution Detection

An open-world learner should be capable of not only discriminating between known classes but also identifying and rejecting unknown classes, which are often cast as Out-of-Distribution (OOD) samples. As identified by [53], earlier attempts towards OOD detection can be grouped into two major categories: (i) Inference methods that use an acceptance score function and (ii) Feature space regularization methods that either alter the network architecture or how it is trained to improve model's robustness to outlier inputs.

Some of the common scoring functions for the first category are based on thresholding the output layers [28, 37, 56], training one-class networks [18, 50] or distance-based scoring [36, 7, 54]. While approaches that threshold the network scores have been popular, they have an underlying requirement, the ability of the network to classify the known classes while providing a scoring mechanism that may be used for that approach. For approaches of Hendrycks and Gimpel [28], and ODIN [38], softmax scores for the known classes are required. While [28] directly thresholds them, [38] incorporates temperature scaling to adjust the softmax output. As witnessed in Fig 1(b), if a softmax classifier is trained on a self-supervised features space it ends up classifying all out-of-label samples as knowns with high probabilities. This would make thresholding approaches impractical for an open world learning problem where the ability to identify unknowns is must. Lee *et al*. [36] employed an acceptance score function by learning a linear classifier to combine the class-conditional Mahalanobis distance metric across multiple CNN layers. [18], and [50] adapted maximum-margin one-class networks to learn features that enable anomaly detection. A variant of distance based classifiers was provided by [7, 54] where they used weibull distributions on the distances between samples in feature space to create an unknowness detector. Our approach employs their techniques, and as visible in Fig 1(c) the distance based techniques work much better than the rest.

The second category of OOD detection methods alters the network architecture or training by (i) training one-vs-rest classifiers [18, 50], (ii) introducing background class regularization [16] or (iii) relying on generative models [46]. Such methods generally need labeled data that is known to not be belonging to any of the known classes. Since labeled data is a scarcity in the open world and more over if a data that was earlier learnt as not belonging to any of the known classes needs to be learnt as a new class, it will require major network training which can destroy the online learning concept of open world learning.

Detecting unknown samples as outliers is an essential component of our Open-World framework. However, as previously mentioned, the unknown in an Open-World setting may be an Out-of-Label-Space sample or an Out-of-Distribution sample depending upon the world of the underlying feature extractor and the unknown sample. We evaluate the unknown detector of our Open-World Model under both of these distinct problem domains.

2.3. Open-World Learning Related Research

Novel Class Discovery and Recognition Recently some aspects of open-world learning have also been studied under the title of *novel class discovery and recognition e.g*., [23, 22, 9]. In [23] authors studied the problem of discovering new classes in unlabeled data via clustering while leveraging the knowledge from labeled data to improve the quality of clustering and to estimate the number of classes in the unlabeled partition. This approach operates in an artificial setting where a partition of labeled and unlabeled data sets is pre-assumed and lacks the capability to isolate labeled samples from unlabeled samples at test time. Unlike true open-world learning setting where new classes are discovered and learned in each incremental phase for several iterations, this approach only discovers and recognize new classes only once and without the capability of unknown detection. Their extended work [22] is relatively closer to our approach where they also first learn feature representation using both labeled and unlabeled data under a self-supervised auxiliary task *i.e*. rotation prediction [20]. A classification head equal to the number of classes in the labeled set along with a softmax layer is then added. The last macro block of the CNN and the classification head is then fine-tuned using labeled data. To discriminate between different classes in the unlabeled set, rank statistics is employed to generate pseudo-labels and subsequently a different classification head for the unlabeled classes is added to the feature extractor. The network is then fine-tuned with these two classification heads by jointly optimizing the two objectives specifically based on labeled and unlabeled partitions. They further explore a setting analogous to incremental learning where the classification head is extended to be equal to the number of classes in labeled and unlabeled set and cross-entropy part of the loss is evaluated on both labeled and unlabeled data. It should be noted in both works [23, 22], the partition of labeled and unlabeled sets is largely unnatural *e.g*., in case of ImageNet experiments, 882 classes belong to labeled set whereas only 30 classes belong to unlabeled set. Further in [22], taking ImageNet experiments as an example, after self-supervised representation learning, the network is first trained on labeled data for 100 epochs, then using joint data for 90 epochs and for incremental setting for 150 epochs.

The concurrent work presented in [9] is based on the similar principles as that of [22] where a feature embedding is learned jointly based on labeled and unlabeled data sets and subsequently linear classification heads are added based on the ground truth number of classes in the labeled set and the expected number of classes in the unlabeled set. The last layers in the CNN along with the classification heads are then fine-tuned using compound objective based on classification, pairwise similarity and regularization. The major contribution of this approach is the introduction of the uncertainty based adaptive margin in the supervised objective that mitigates the bias towards the seen classes. Differently from [22], they operate on a more realistic setting where the number of classes belonging to labeled and unlabeled partitions are more balanced *e.g*. in case of ImageNet, 50 classes belong to the seen/labeled partition whereas another 50 are considered novel classes in the unlabeled partition. The feature extractor in this approach is also based on self-supervised learning, however they rely on SimCLR [14] instead of RotNet [20]. Same as [23, 22], this approach also conducts discovery and recognition of unknown classes just once.

Open-Set Recognition While none of the previous works have focused on true open world learning, there has been prior work which was very closely related [6, 54]. Both the works focused on the problem of Open-Set incremental learning rather than open world learning. This means that rather than attempting to find unknowns and then learn them as new classes, these works tried to learn new classes in a supervised manner while identifying some of the samples as unknowns. Furthermore, [54] used a flawed protocol where their unknown classes that were being identified and incrementally learnt were a subset of the classes that were used to train their supervised feature space. This provided a good inherent feature separation between their known classes and the supposed “unknown” classes while leading to unrealistically promising good results.

These issues have inspired us to propose a true open-world learning protocol (Sec 4) to help progress the future research directions.

3. Our Approach

In this section, we first discuss the current trends of research in incremental learning and how our approach is well-suited for the problems of incremental and open-world learning. Next, we detail our incremental approach followed by our open-world learning approach.

Data availability vs label availability While incremental learning was initially considered an important research problem because of lack of availability of training data, with recent advances in self supervised learning research, it has become more apparent that the problem isn’t lack of availability of training data but rather lack of availability of labeled training data. While incremental learning cannot directly benefit from the self-supervised learning techniques due to their high training times which may lead to degeneration of the timeliness component for the incremental learning problem, they can definitely use the advances in self supervised learning to initialize their learning feature space. Our approach capitalizes on this component by initializing our feature space with features trained using self supervised learning techniques. Recently, there have been many attempts towards learning better feature representations under self-supervised learning by defining a *pretext task* [47, 49, 59, 60, 35, 20, 21, 45, 48, 31, 44, 33, 19], using *contrastive loss* [27, 14] or by *clustering* deep features [10, 11]. As demonstrated later in Sec 5, our underlying feature representations can be augmented from any of these advances in self-supervised learning domain.

Learning without Back-propagation Most recent approaches towards incremental learning are based on retraining the networks with the newly added classes and keeping the exemplars of the old classes to circumvent catastrophic-forgetting. At every incremental stage, multiple rounds of back-propagation are essential for good performance of such methods and hence require offline training. Unlike these approaches, our method is based on Extreme Value Machines (EVMs) [54] where we fit a single EVM based on multiple Weibull distributions per class. Since, our incremental learning classifier is not based on altering the feature representation and does not require back-propagation, we can operate in a true **online** manner and do not require rounds of offline training like other approaches. Since, our model is not sharing its learning capacity between old and new classes, rather recognition capacity is enhanced in the form of new EVMs as new classes are added, it does not suffer from catastrophic forgetting.

Learning with deeper networks Supervised learning domain has long benefited from advances in neural network architectures which generally involve a higher number of training parameters. Unfortunately, the current incremental learning approaches are stuck with shallower networks like ResNet-18, because the small amount of data (even in initial incremental phase) available to them during training might result in network over-fitting. Since incremental learning approaches have to also rapidly adapt to new classes, any incremental learning process involving learning a deeper network becomes expensive. Unlike supervised learning approaches that are dependent on labeled data, self supervised approaches thrive on the abundance of unlabeled data. We leverage this capability of self-supervision which helps us also to utilize network architectures with a higher number of parameters, without running into issues such as over-fitting. Since our approach does not involve back propagation it also makes the process to learn new classes faster.

Learning without Exemplars Additionally, the EVM for new classes can be fitted without retaining the exemplars from the old-classes which corresponds to being a **zero-exemplar** model. We also show that our approach is capable of performing even better when we include exemplars while fitting the new EVMs. As the EVM framework is based on retaining the number of extreme vectors which are part of our representation model, the exemplars we use are essentially part of our model and not explicit images or feature vectors separately maintained in additional memory, so variant of our approach leveraging the extreme vectors is still a zero-exemplar model. It should be noted that earlier methods have specific memory budget, *e.g.*, 20 exemplars/class to retain the samples of old classes which is an additional memory requirement than the network weights. Approaches like [30] try to retain feature vectors instead of images then requiring additional adaptation network to project old features to new feature space and hence additional training of the secondary network is also required. Whereas, our approach does not require additional memory to maintain these exemplars as the exemplars (extreme vectors) being used are part of our underlying EVM model. As demonstrated in [17] (Tab 4), as the number of exemplars per class decreases the performance for all the approaches decreases which is not the case in our approach.

NIL – Non-Backpropagating Incremental Learning Our incremental learning approach, NIL belongs to the first category of the various incremental learning approaches discussed in section Sec 2.1 *i.e.* our feature representation is fixed. We use self supervised learning techniques to learn a feature space from unlabeled data that can be used for incrementally learning new classes. At the zero-th stage of our incremental learning framework, the class labels for initial C classes are provided and an EVM [54] is fitted for each of these classes. The EVM model for each class is comprised of multiple Weibull distributions (Φ_i), where the probability of a sample x' belonging to the distribution may be found with $\Psi\left(x_i, x'; \Phi_i\right)$ where x_i is the underlying feature vector for i-th instance of a class which has been maintained as representation of that specific class and called as an extreme vector. Please see [54] for more details. At each subsequent incremental step, K new classes are introduced and new EVM models are fitted for these new classes. The newly introduced K EVMs are appended to the existing EVM model for C classes and the process repeats.

NOWL – Non-backpropagting Open World Learning The Open-World works in operational phases; each of which is comprised of an uncontrolled real world enviornment where an agent finds unknowns and then learns them in an incremental learning stage. Our Open-World Model NOWL is first learned with *C* classes for which the labels are available. We use the Extreme-Value-Machine (EVM) for detecting the unknown (out-of-distribution or out-of-label space) samples and NIL for learning them in future. Once, the initial model is learned, it operates in the operational phase where it performs C class classification as well as unknown detection. The probability that a query point x' (in the form of feature vector) is associated with a class $\mathcal{C}_l$ is $\hat{P}(\mathcal{C}_l|x') = \arg\max_{\{i:y_i=\mathcal{C}_l\}} \Psi\left(x_i, x'; \Phi_i\right)$. Given $\hat{P}$, we can determine whether x' belongs to an existing known class or is an unknown using:

$$y^* = \begin{cases} \arg\max_{l\in\{1,\ldots,M\}} \hat{P}(\mathcal{C}_l|x') & \text{if } \hat{P}(\mathcal{C}_l|x') \geq \delta \\ \text{“unknown”} & \text{Otherwise.} \end{cases} \tag{1}$$

After enrolling the newly detected and labeled k_1 classes through incremental steps, the OWM model is now capable to perform multi-class classification for $C + k_1$ classes while detecting the subsequent unknowns.

4. Open-World Learning Protocol

As discussed in Sec 2, open-world learning can be broadly categorized into two sub-domains, out-of-distribution detection and incremental learning. In this section we bridge the gap between the protocols followed for these two independent research domains by proposing our open-world learning protocol (Fig 3) which can help advance research in this area of *prime practical importance*. Any approach addressing the open-world problem has to identify unknown samples, get them labeled by an annotator (human or autonomous agent) and learn them for future encounters. The most important distinction between an incremental learning problem and an open-world learning problem lies in temporal encounters with unknown data, when compared to out-of-distribution research the major difference is the continuous change in the data labels between each phase from unknowns to knowns. In order to mimic these real-world temporal chain of events, the protocol presents the data to the algorithm in a set of phases which can be broadly classified into three phases for easier understanding.

Initialization Phase ω_0 Before entering the real-world operational phases the algorithm needs to go through an ini-

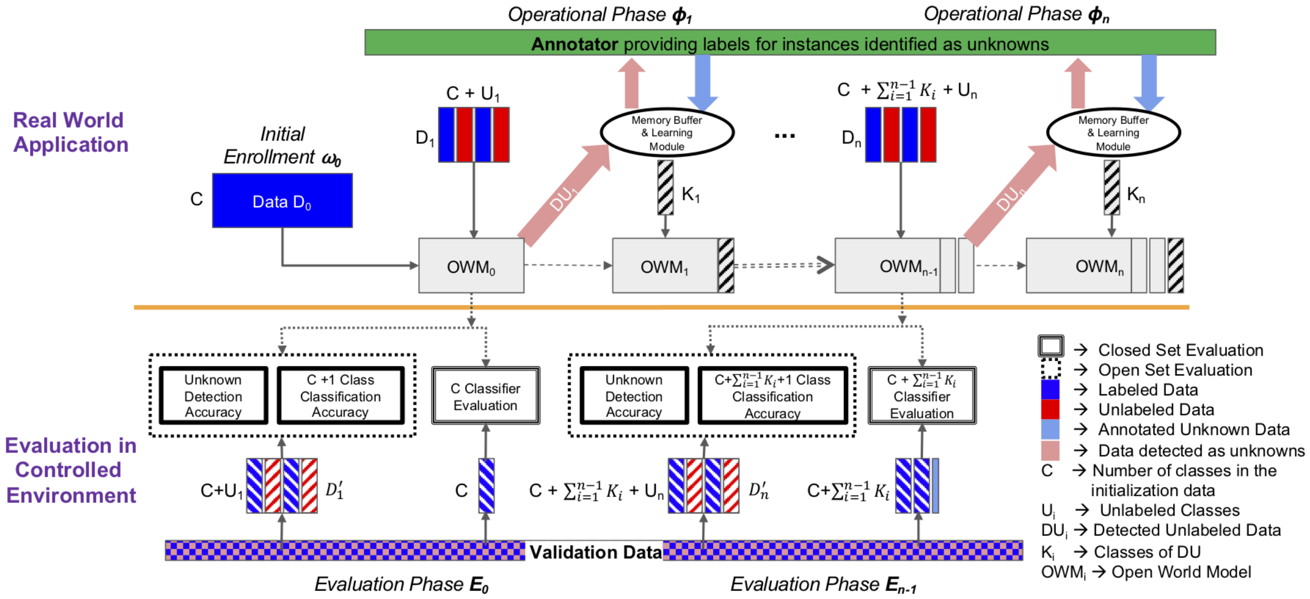

Operational Phase $\boldsymbol{\phi_1}$
Operational Phase $\boldsymbol{\phi_n}$
Annotator providing labels for instances identified as unknowns
Real World Application
Initial Enrollment $\boldsymbol{\omega_0}$
C
Data D_0
C + U_1
D_1
Memory Buffer & Learning Module
DU_1
K_1
...
C + $\sum_{i=1}^{n-1} K_i$ + U_n
D_n
Memory Buffer & Learning Module
DU_n
K_n
OWM_0
OWM_1
OWM_{n-1}
OWM_n
Unknown Detection Accuracy
C +1 Class Classification Accuracy
C Classifier Evaluation
Unknown Detection Accuracy
C+$\sum_{i=1}^{n-1} K_i$+1 Class Classification Accuracy
C + $\sum_{i=1}^{n-1} K_i$ Classifier Evaluation
Evaluation in Controlled Environment
C+U_1
D_1'
C
C + $\sum_{i=1}^{n-1} K_i$ + U_n
D_n'
C+$\sum_{i=1}^{n-1} K_i$
Validation Data
Evaluation Phase $\boldsymbol{E_0}$
Evaluation Phase $\boldsymbol{E_{n-1}}$
→ Closed Set Evaluation
→ Open Set Evaluation
🟦 → Labeled Data
🟥 → Unlabeled Data
→ Annotated Unknown Data
→ Data detected as unknowns
C → Number of classes in the initialization data
U_i → Unlabeled Classes
DU_i → Detected Unlabeled Data
K_i → Classes of DU
OWM_i → Open World Model

Fig. 3: OPEN-WORLD LEARNING *PROTOCOL: Initially the Open-World Model OWM is trained with data belonging to C classes. Afterwards, in each phase ϕ_n; the OWM operates in a conventional open-world setting and then learns incrementally. In open-world setting, OWM performs multi-class classification on the seen classes while also detecting and accumulating unlabeled data belonging to new classes in a memory buffer. An annotator then provides the K_n class labels for the detected unlabeled samples (DU_n) present in the memory buffer. The OWM model is then updated for these new K_n classes and is now capable of multi-class classification for $C + \sum_{i=1}^{n} K_i$ classes. In each phase, using the validation set, we evaluate our OWM model first for open-world performance i.e. multi-class classification and unknown detection and then on incremental learning task.*

tialization phase (ω_0). In this phase the protocol provides labeled data D_0 from which the algorithm learns to classify C classes. The resulting open-world model, identified as OWM_0, is now ready to be deployed in the real-world scenario.

Operational Phase ϕ_n This phase mimics the real-world operation where an agent is responsible for classifying the data into classes which it has already learnt as well as identifying data belonging to unknown classes. Here the algorithm operates in phases with some feedback from the human annotator which is simulated in this protocol. In operating phase ϕ_1 the protocol presents the algorithm with samples not just from the C classes (learnt in ω_0) but also from unlabeled/unknown classes U_1. At this point the model OWM_0 is responsible for identifying unknown samples from D_1 as DU_1 (detected unknowns) while also correctly classifying any of the samples in D_1 that belong to C. The detected unknowns DU_1 then depending on the algorithm may be held in a memory buffer before being sent to an annotator for obtaining labels. Once the annotator provides labels for all samples in DU_1, the algorithm now has labeled samples from K_1 classes, where $K_1 \subset U_1$. Using these samples, the algorithm is able to adapt itself to classify classes $C+K_1$. This protocol continues for various such operational phases ϕ_n, where in each step, the classes in D_n increase as $C + \sum_{i=1}^{n-1} K_i + U_n$.

Evaluation Phase E_n This phase is aimed at evaluating various aspects of the agent's performance. We use three major metrics to report a model's performance under the open-world and the closed-world setup. ***(a) Closed-world Classification Accuracy (CCA)*** The closed-world classification accuracy for any given evaluation phase E_n is measured as the accuracy in terms of the samples belonging to the classes $C + \sum_{i=1}^{n-1} K_i$. ***(b) Open-world Performance*** In order to measure a model's open-world performance at phase E_{n-1}, we feed the model samples from a validation set D_n' that contains samples from $C + \sum_{i=1}^{n-1} K_i + U_n$. We break down the open-world performance of model OWM_{n-1} in the following two parts. ***(i)*** *Unknown Detection Accuracy (UDA)* This measure provides us an insight towards the algorithm's ability to detect unknowns. It is simply the binary class accuracy for identifying the samples as unknowns. ***(ii)*** *Open-world Classification Accuracy (OCA)* The open–world classification accuracy combines the CCA and UDA by considering the problem to be classifying the sample into Knowns+1 classes.

Protocol Variants Based on the above descriptions we propose OpenWorld (OW)-100 and OW-500 protocols. These protocols use images from ImageNet 2012 [55]. OW-100 protocol uses the same classes as in ImageNet-100 incremental protocol [52], while OW-500 keeps extending OW-100 to 500 classes with more batches. Since the open-world protocols use unknowns in order to test a model's performance, the algorithms have to follow a basic constraint

Tab. 1: Average incremental Top-1 accuracy, NIL vs state of the art: *Even while using* ***zero-exemplars****, NIL outperforms PODNet [17] (SOTA) on 10 class increments and is comparable to its performance on 5 class increments. Additionally, when using exemplars equivalent to all other approaches here, NIL out performs [17] across all incremental steps* ***without any back propagation learning****. Results with * are from [29], and with* † *generated by [17]. Results with* ‡ *are from [39] where the original approaches were combined with the mnemonics framework.*

	50 steps	25 steps	10 steps	5 steps
New classes per step	**1**	**2**	**5**	**10**
iCaRL* [52]	—	—	59.53	65.04
iCaRL† [52]	54.97	54.56	60.90	65.56
iCaRL‡ [52]	—	67.12	70.50	72.34
BiC† [58]	46.49	59.65	65.14	68.97
BiC‡ [58]	—	69.22	70.73	71.92
UCIR (CNN)* [29]	—	—	68.09	70.47
UCIR (CNN)† [29]	57.25	62.94	67.82	71.04
UCIR (CNN)‡ [29]	—	69.74	71.37	72.58
PODNet (CNN)† [17]	62.48	68.31	74.33	75.54
NIL	**66.49**	**72.36**	**76.33**	**78.03**
NIL (Zero Exemplar)	—	58.78	74.22	**77.42**

regarding the training data which is also commonly followed in the out-of-distribution detection research. The model's underlying feature representations should not be explicitly trained to classify any of the classes in ImageNet 2012 which may be used as unknowns in the protocol.

Protocol Extension and Adaptability Although described in a supervised setting, the proposed protocol is applicable and extendable to semi-supervised and unsupervised open-world settings as well. For example, instead of getting annotations from a human annotator for all the discovered unknowns, a subset of unknown samples can be annotated. The detected unknowns can be first clustered and then only representative samples per cluster (*e.g.*, cluster centroids) are annotated by the human operator. Similarly human-in-the-loop setting can be avoided altogether and instead pseudo-labels can be generated by an autonomous annotator *i.e.* a clustering method. To this end, such a formulation can benefit further from approaches that estimate the number of clusters or improve clustering quality by leveraging knowledge from labeled/known data *e.g.* [23]. We intend to explore these dimensions of open-world learning and variants of our protocol in future work.

5. Experimental Results

Our experiments and their results are comprised of the following two parts.

Incremental Learning Our comparative results for class incremental learning module are documented in Tab 1 where we compare our approach against current state-of-the-art (SOTA) in incremental learning [52, 13, 58, 17, 39]. We follow iCaRL [52] ImageNet-100 protocol which starts with initially learning 50 classes and then adds classes incrementally. While the original protocol added 10 and 5 classes per incremental step, [17] explored smaller incremental steps of introducing 2 and 1 class/es at each step. We report the average incremental Top-1 accuracy on all of the above. We compare our approach to the most recent work [17] on incremental learning and additionally from the *mnemonics* approach [39]. The numbers reported for each of the earlier approaches are based on 20 exemplars per class. For our approach, we report numbers for both cases *i.e.*, when the EVM models for new classes are added without information from the old class (zero exemplar) and when a chosen number of extreme vectors per class (20) serve as the negative examples for the new classes. As is apparent from Tab 1, even with our **zero-exemplar** model, we are capable of outperforming the latest SOTA on 10 class increments or comparable to its performance on 5 class increments. Additionally, when the new EVMs [54] are fitted using 20 extreme vectors from each of the old-classes as negatives, we outperform [17] across all incremental steps. An underlying limitation of our zero-exemplar formulation is that there should be at least two classes per incremental step as the EVM model for one class is fitted with respect to the negatives taken from the other class/es. This reflects in second-last row of Tab 1, as the number of classes per incremental step increases, the performance increases. Our experiments with NIL utilize features from SwAV [12], trained with unlabeled ImageNet data for the ResNet-50 model. Details about the EVM parameters may be found in the supplemental material. As discussed in Sec 3, both our approaches, NIL and NOWL , are orthogonal to the current advances in self-supervised learning. Tab 2 demonstrates that NIL can be augmented with any of the self-supervised feature representation approaches. For each algorithm, we report their Top-1 accuracy for each of the underlying feature extractor which is kept frozen and a linear classifier layer is learnt in a supervised manner. It is interesting to note that the average class incremental accuracy is in accordance to that of the supervised Top-1 accuracy i.e., for a poor performer on Top-1 supervised accuracy (*e.g.*, MOCOv1), the incremental performance is also poor and vice-versa.

Open World Learning For our open world learning experiments we use the OpenWorld-100 (OW-100) and OpenWorld-500 (OW-500) protocols proposed in Sec 4. We summarize the results in Tab 3. All our experiments involve a MoCov2 with ResNet-50 backbone. Based on the unlabeled data that was used for training the self-supervised learning

Tab. 2: NIL AUGMENTED WITH VARIOUS SELF-SUPERVISED TECHNIQUES *Average incremental Top-1 accuracy for varying self-supervised techniques with the ImageNet-100 protocol*

Approach	***Top-1 Acc***	***New Classes per Step*** *1*	*2*	*5*	*10*
MOCOv1 [27]	60.6%	53.54	56.45	59.43	61.94
SimCLR 1x [14]	69.1%	60.39	64.51	67.78	69.15
MOCOv2 [15]	71.1%	**66.96**	69.90	74.18	75.11
SeLa-v2 [1]	71.8%	53.60	67.80	71.58	71.90
DeepCluster-v2 [12]	75.2%	63.70	71.18	75.35	76.33
SwAV [12]	**75.3%**	66.49	**72.36**	**76.33**	**78.03**

Tab. 3: OPENWORLD PERFORMANCE FOR NOWL *Below we report the performance of the our NOWL approach on the proposed OpenWorld-100 (OW-100) and OpenWorld-500 (OW-500) protocols. Based on the unlabeled data used during training of these networks we report Out-Of-Label space (OOL) and Out-Of-Distribution (OOD). The reported numbers are average accuracies across batches, for more results and performance details on each batch/step, please refer to the supplemental material.*

Protocol Details	**# Ex-em-plars**	**New classes per step $\|U_n\|$**					
		5			10		
		# Knowns/ # Unknowns *2500/1250*			*# Knowns/ # Unknowns* *2500/2500*		
		UDA	OCA	CCA	UDA	OCA	CCA
OW-100 (OOL)	0	65.16	55.56	69.83	66.53	55.62	72.57
	20	72.69	55.25	72.94	68.43	55.90	73.44
OW-100 (OOD) OpenImages	0	59.02	35.72	50.08	62.43	37.25	54.33
	20	68.33	36.2	55.24	65.6	37.35	56.0
OW-100 (OOD) Places	0	54.84	30.0	42.39	60.0	31.87	46.49
	20	67.56	31.24	48.24	65.33	32.75	49.6
OW-500 (OOL)	20	70.11	45.65	56.96	68.54	45.95	57.77

approach we make the distinction of an Out-of-Label-space (OOL) experiment versus and Out-of-Distribution(OOD) experiment. For our **OOL** experiments we trained the MoCo-v2 with unlabeled ImageNet 2012 data. For the **OOD** experiments we further provide two variants. In one case we use the unlabeled Places2 [62] as the training dataset and in the other we use the OpenImages-v6 [34] as the training dataset. From the Tab 3 we observe that it definitely helps improve the classification accuracy if the network has been initialized with data that is closer to the testing world (Fig 2). For example, OpenImages data is closer to the ImageNet data as compared to the Places2 data which has less resemblance to ImageNet. For our unknowness probability threshold we use a score of 0.5, more details regarding the EVM parameters can be found in our supplemental material.

6. Discussion & Conclusion

Open-world learning strongly depends on the ability to both detect unknown samples and incrementally learn from them once detected. This paper shows that using self-supervised features improves the performance of both tasks. Our experiments with non-back-propagating incremental learning (NIL) showed that using appropriate self-supervised feature spaces improves incremental learning under the standard ImageNet-100 incremental learning protocol. Using self-supervised feature space allowed our EVM-based approach to advance the state-of-the-art on incremental learning using 20 exemplars per classes. We investigated multiple self supervised features and showed that all of the self-supervised feature spaces allowed incremental learning without forgetting even when using zero-exemplars – *i.e.*, without any update to the feature space or model for the prior classes. This allows straight-forward edge-learning of new classes as access to any prior class examples is not required. Our incremental approach well satisfies the desired properties of such methods as defined in [5].

Our paper formalizes the novel concept of Out-of-Label (OOL) space detection and hypothesizes it is easier than Out-of-Distribution (OOD) detection due to feature space over-fitting and hence OOD projection overlap inherent to supervised feature-spaces. We argue that self-supervised data with mostly OOL learning is a more natural setting for open-world learning because collecting data is easy, while labeling data is more difficult. Our NOWL experiments provide new state-of-the-art results for open-world learning in both OOL and OOD settings and studied the impact of different self-supervised feature spaces, including ImageNet space, moderately related OpenImages-v6 [34], and a less related Places2 [62] space. These experiments validate our hypothesis that OOL is easier than OOD, at least for the Extreme-Value-Theory based classifiers and considered self-supervised sub-spaces. We conjecture it would be true for any feature space and classifier, but further work with a broader range of classifiers is needed to confirm that conjecture.

Readers might note that experiments for NOWL lack comparison with any other algorithms. This is because we are building on top of the state-of-the-art EVM [54], demonstrating that the choice of feature space is critical. We did have to adapt the EVM to support incremental learning with zero or 20 exemplars rather than always using all data. One might also note that in the NOWL experiments, we used MoCov2 despite SaWV being better in the NIL experiments. This is due to the availability of MocoV2 pre-trained networks from Places2 [62] and OpenImages-v6 [34], and paper's focus being on understanding the problem and not absolutely maximizing the performance.

One of the advantages of using self-supervised feature spaces is the amount of data available, which in turn allows for deeper/better networks. We note that traditional incremental feature learning has, to date, been using smaller networks like ResNet-18; research is lacking on how well such small network work with large number of classes or how incremental systems would transition to employing larger networks during the incremental steps. Herein we used ResNet-50 for our models, but we could have easily used even larger networks, which would likely improve performance but this investigation is left for future work since it is not directly related to the core hypotheses of this paper.

While the proposed approach has many advantages and advanced the state-of-the-art, there are still some important limitations. First, there is range of system issues yet to be studied that could improve performance and provide new insights. These include the bigger/better networks, intelligent selection of exemplars, using Weibull probabilities during class updates and a hybrid feature space that addresses domain adaption as it learns new classes but also retains the base self-supervised space.

Open-world learning is a growing research sub-field leveraging feature learning, incremental learning, and out-of-distribution/out-of-label-space detection. This paper introduced the first open world protocol and its evaluation paradigm which builds on top of the widely used incremental learning protocol.

Acknowledgement

This research is sponsored by the Defense Advanced Research Projects Agency (DARPA) under HR001120C0055. The views contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA or the U.S. Government.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Int. Conf. Learn. Represent.*, 2020. 9

[2] Eden Belouadah and Adrian Popescu. Deesil: Deep-shallow incremental learning. In *Eur. Conf. Comput. Vis. Work.*, 2018. 3, 4

[3] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Int. Conf. Comput. Vis.*, 2019. 4

[4] Eden Belouadah and Adrian Popescu. Scail: Classifier weights scaling for class incremental learning. In *IEEE Win. Conf. App. Comput. Vis.*, 2020. 4

[5] Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *ArXiv*, abs/2011.01844, 2020. 4, 9

[6] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1, 5

[7] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 4

[8] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249 – 259, 2018. 4

[9] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *ArXiv*, arXiv:2102.03526, 2021. 4, 5

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Eur. Conf. Comput. Vis.*, 2018. 5

[11] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Int. Conf. Comput. Vis.*, 2019. 5

[12] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Adv. Neural Inform. Process. Syst.*, 2020. 8, 9

[13] Francisco M. Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Eur. Conf. Comput. Vis.*, 2018. 8

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. on Mach. Learning*, 2020. 5, 9

[15] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *ArXiv*, 2010.15277, 2020. 9

[16] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Adv. Neural Inform. Process. Syst.*, 2018. 1, 4

[17] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Eur. Conf. Comput. Vis.*, 2020. 1, 3, 6, 8

[18] Sarah M. Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 52(October), 2016. 1, 4

[19] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Eur. Conf. Comput. Vis.*, 2020. 1, 5

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Int. Conf. Learn. Represent.*, 2018. 4, 5

[21] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Int. Conf. Comput. Vis.*, 2019. 5

[22] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Int. Conf. Learn. Represent.*, 2020. 1, 4, 5

[23] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Int. Conf. Comput. Vis.*, 2019. 4, 5, 8

[24] Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural net-

work to prevent catastrophic forgetting. In *Eur. Conf. Comput. Vis.*, 2020. 4

[25] Tyler L. Hayes and Christopher Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2020. 3, 4

[26] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 5, 9

[28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Int. Conf. Learn. Represent.*, 2017. 1, 4

[29] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 3, 4, 8

[30] Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In *Eur. Conf. Comput. Vis.*, 2020. 1, 3, 6

[31] Simon Jenni, Hailin Jin, and Paolo Favaro. Steering self-supervised feature learning beyond local pixel statistics. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 5

[32] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019. 2

[33] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *PAMI*, 2020. 1, 5

[34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.*, 128:1956–1981, 2020. 9

[35] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 5

[36] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Adv. Neural Inform. Process. Syst.*, 2018. 1, 4

[37] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018. 1, 3, 4

[38] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Int. Conf. Learn. Represent.*, 2018. 1, 4

[39] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 3, 8

[40] Yong Luo, Liancheng Yin, Wenchao Bai, and Keming Mao. An appraisal of incremental learning methods. 22(11), 2020. 4

[41] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation. *ArXiv*, 2010.15277, 2020. 4

[42] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *Eur. Conf. Comput. Vis.*, 2012. 2

[43] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2624–2637, 2013. 2

[44] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 5

[45] T. Nathan Mundhenk, Daniel Ho, and Barry Y. Chen. Improvements to context based self-supervised learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 5

[46] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *Int. Conf. Learn. Represent.*, 2019. 4

[47] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Eur. Conf. Comput. Vis.*, 2016. 5

[48] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Int. Conf. Comput. Vis.*, 2017. 5

[49] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 5

[50] Pramuditha Perera and Vishal M. Patel. Learning deep features for one-class classification. *IEEE Trans. Image Process.*, 28(11):5450–5463, 2019. 1, 4

[51] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. itaml: An incremental task-agnostic meta-learning approach. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 4

[52] Sylverstre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifiers and representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 3, 4, 7, 8

[53] Ryne Roady, Tyler L Hayes, Ronald Kemker, Ayesha Gonzales, and Christopher Kanan. Are open set classification methods effective on large-scale datasets? *PLOS ONE*, 15(9), 2020. 1, 4

[54] Ethan M Rudd, Lalit P Jain, Walter J Scheirer, and Terrance E Boult. The extreme value machine. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):762–768, 2017. 1, 4, 5, 6, 8, 9

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 2, 7

[56] Lei Shu, Hu Xu, and Bing Liu. DOC: Deep open classification of text documents. In *Empirical Methods in Natural Language*, 2017. 1, 4

[57] Max Welling. Herding dynamical weights to learn. In *Int. Conf. on Mach. Learning*, 2009. 3, 4

[58] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 3, 4, 8

[59] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Eur. Conf. Comput. Vis.*, 2016. 5

[60] Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 5

[61] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shutao Xia. Maintaining discrimination and fairness in class incremental learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 4

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018. 2, 9