Out-Of-Distribution Detection In Unsupervised Continual Learning

Jiangpeng He
he416@purdue.edu

Fengqing Zhu
zhu0@purdue.edu

Elmore Family School of Electrical and Computer Engineering, Purdue University, USA

Abstract

Unsupervised continual learning aims to learn new tasks incrementally without requiring human annotations. However, most existing methods, especially those targeted on image classification, only work in a simplified scenario by assuming all new data belong to new tasks, which is not realistic if the class labels are not provided. Therefore, to perform unsupervised continual learning in real life applications, an out-of-distribution detector is required at beginning to identify whether each new data corresponds to a new task or already learned tasks, which still remains under-explored yet. In this work, we formulate the problem for Out-of-distribution Detection in Unsupervised Continual Learning (OOD-UCL) with the corresponding evaluation protocol. In addition, we propose a novel OOD detection method by correcting the output bias at first and then enhancing the output confidence for in-distribution data based on task discriminativeness, which can be applied directly without modifying the learning procedures and objectives of continual learning. Our method is evaluated on CIFAR-100 dataset by following the proposed evaluation protocol and we show improved performance compared with existing OOD detection methods under the unsupervised continual learning scenario.

1. Introduction

Unsupervised continual learning is an emerging future learning system, capable of learning new tasks incrementally from unlabeled data. It requires neither static datasets nor human annotations compared with supervised offline learning. Existing methods study this problem under the assumption that all new data belongs to new tasks. We argue that if human annotation is not available as common in unsupervised scenario, we cannot know whether the unlabeled new data belongs to new or learned tasks. For example, an image-based mobile food recognition system should be able to distinguish new and learned food images first instead of blindly treating all of them as new food classes to perform unsupervised continual learning for update. Therefore, in order to make unsupervised continual learning work in practical problems, an out-of-distribution (OOD) detector should be required at the beginning of each incremental learning step to identify whether each data belongs to new or already learned tasks. However, the problem of OOD detection in continual learning still remains under-explored, *i.e.* none of the existing OOD detection methods target for continual learning.

The goal of OOD detection for image classification is to detect novel classes data. However, it becomes more challenging under continual learning scenario due to (1) the training data of learned tasks becomes unavailable; (2) we also need to address catastrophic forgetting problem [18]. Most existing methods cannot be applied here because they either require all training data for already learned tasks to train an OOD detector [13, 25, 28], or they need to modify the training procedure and objectives [9, 19, 20, 29], which may sacrifice the classification accuracy. Therefore, we focus on "post-hoc" methods [31] that can be directly applied on any trained classification models to perform OOD detection based on the output confidence, which has been widely adopted in real-world environments to avoid the need to access training data.

The central idea of "post-hoc" methods to perform OOD detection is to assign in-distribution (ID) data with higher confidence value $Conf_{in}$ than the OOD data $Conf_{out}$ based on the output vector where the confidence $Conf$ is defined as the maximum of softmax output [10, 15] or the energy score [16]. The detection performance greatly depends on the difference value of output confidence between ID and OOD data $\mathcal{D}_c = Conf_{in} - Conf_{out}$ where higher $\mathcal{D}_c$ indicates better discrimination. However, there exists two major issues in continual learning scenario that can lead to the decrease of $\mathcal{D}_c$ including (1) the biased output value towards new classes as revealed in [30, 33]; (2) the decrease of output confidence compared with offline learning due to the objective of improving generalization ability to mitigate catastrophic forgetting [14, 32]. Both issues can result in performance degradation for existing "post-hoc" methods.

In this work, we first formulate the OOD detection in unsupervised continual learning scenario denoted as OOD-UCL and introduce the corresponding evaluation protocol. Then, we propose a novel OOD detection method that can address both issues mentioned above to achieve improved performance in unsupervised continual learning scenario. The main contributions are summarized as following.

- To the best of our knowledge, we are the first to formulate the problem and evaluation protocol for out-of-distribution detection in unsupervised continual learning (OOD-UCL), which remains under-explored.
- A novel method is introduced for OOD detection by correcting output bias and enhancing output confidence difference based on task discriminativeness.
- We conduct extensive experiments on the CIFAR-100 [12] dataset to show the effectiveness of each component of our proposed method compared to existing works under OOD-UCL scenario.

2. Related Work

We focus on image classification problem and we review the existing methods that are related to our work including (1) unsupervised continual learning; (2) OOD detection.

2.1. Unsupervised Continual Learning

Compared with supervised case, unsupervised continual learning has not received much attention [17]. Stojanov *et al.* [27] introduced an unsupervised object learning environment to learn a sequence of single-class exposures. In addition, CURL [22] and STAM [26] are proposed for task-free unsupervised continual learning where task boundary is not given. Based on existing supervised protocol [23], the most recent work [5] proposed to use pseudo labels obtained based on cluster assignments to perform continual learning and show promising results on several benchmark datasets in unsupervised scenario. However, they only assume a simplified scenario where all the new data belong to new classes, which rarely happens in real life applications when the class labels are not available. Therefore, an OOD detector that can work under unsupervised continual learning scenario becomes indispensable.

2.2. Out-of-distribution Detection

As illustrated in Section 1, we focus on image classification based OOD detection and analyze this problem in continual learning scenario where the training objective is more challenging. Therefore, we target on methods that can be applied to any trained classification model without modifying the training procedure, which is called "post-hoc" methods [31]. Existing "post-hoc" methods are originated from [10], which directly uses the maximum softmax probability as the confidence score to discriminate ID and

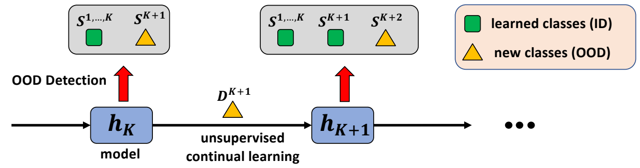

Figure 1. Formulation of out-of-distribution detection in unsupervised continual learning (OOD-UCL). h_K refers to the updated incremental models after learning $\mathcal{T}^K$. D^K and S^K denote the corresponding training and testing splits for task K, respectively.

OOD data. Then ODIN [15] applies temperature scaling and input perturbation to amplify the confidence difference D_c between ID and OOD data where a large temperature transforms the softmax score back to the logit space. Built on these insights, recent work [16] proposed to use energy score as output confidence for OOD detection, which maps the output to a scalar through a convenient log-sum-exp operator. However, none of the existing "post-hoc" methods consider the two issues in continual learning scenario as illustrated in Section 1, resulting in performance degradation.

3. Problem Formulation

The objective is to perform OOD detection in continual learning scenario to discriminate unlabeled learned tasks data (as ID) and new task data (as OOD), which can be then incorporated into any existing unsupervised continual learning methods to apply in real life applications. We formulate the out-of-distribution in unsupervised continual learning (OOD-UCL) problem based on the existing unsupervised class-incremental learning protocol [5] to evaluate the OOD detection performance before each incremental learning step. Specifically, the continual learning for image classification problem $\mathcal{T}$ can be expressed as learning a sequence of N tasks $\{\mathcal{T}^1, ..., \mathcal{T}^N\}$ corresponding to $(N-1)$ incremental learning steps where the learning of the first task $\mathcal{T}^1$ is not included. Each task contains M non-overlapped classes, which is known as incremental step size. Let $\{D^1, ..., D^N\}$ denote the training data and $\{S^1, ..., S^N\}$ denote the testing data for each task, we formulate the OOD-UCL with the following properties.

Property 1: The OOD detection is performed at beginning of the learning step for each new task $\mathcal{T}^K$ where $K \in \{2, ...N\}$. The test data belonging to learned tasks $S^i, i \in \{1, ...K-1\}$ is regarded as ID data and the test data belonging to the current incremental step S^K is regarded as the OOD data. Figure 1 illustrates the evaluation protocol, where we perform total $(N-1)$ times OOD detection for continually learning a sequence of N tasks $\{\mathcal{T}^1, ..., \mathcal{T}^N\}$.

Property 2: The training data allowed for OOD detection before learning $\mathcal{T}^K$ is restricted to (1) the training set of D^{K-1} and (2) the stored exemplars belonging to $\{\mathcal{T}^1, ..., \mathcal{T}^{K-2}\}$ if applicable. This restricts the usage of most existing methods [13, 25, 28] which requires all train-

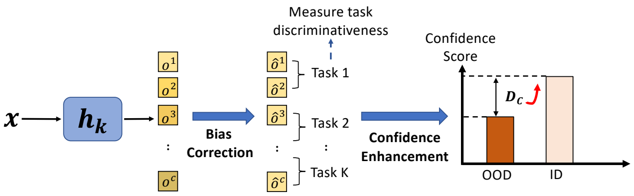

Figure 2. The overview of our proposed method where $\mathbf{x}$ refers to input data and h_K denotes the continual model after learning task $\mathcal{T}^K$. We first correct the bias of output O to obtain $\hat{O}$ and then perform confidence enhancement to further increase the confidence difference D_c to improve OOD detection performance.

ing data for learned classes to train an OOD detector.

Evaluation metrics: In OOD detection, each test data is assigned with a confidence score where samples below the pre-defined confidence threshold are considered as OOD data. By regarding the ID data as positive and OOD data as negative, we can obtain a series of true positives rate (TPR) and false positive rate (FPR) by varying the thresholds. One of the commonly used metrics for OOD detection is **FPR95**, which measures the FPR when the TPR is 0.95 and lower value indicates better detection performance. Besides, we can also calculate the area under receiver operating characteristic curve (**AUROC** [2]) based on FPR and TPR as well as the area under the precision-recall curve (**AUPR** [24]). For both AUROC and AUPR, a higher value indicates better detection performance.

4. Our Method

In this section, we introduce a novel "post-hoc" OOD detection method with the goal of improving the performance under unsupervised continual learning scenario, *i.e.* increase the confidence difference D_c between ID and OOD data for better discrimination. The overview of the proposed method is shown in Figure 2, which can be directly applied without requiring any change to the existing classification models. There are two main steps including **bias correction** and **confidence enhancement** where we first correct the biased output value and then enhance the confidence difference D_c based on task discriminativeness, which are described in Section 4.1 and Section 4.2, respectively.

4.1. Bias Correction

Output bias towards new classes is a widely recognized issue [30, 33] caused by the lack of training data for learned tasks during continual learning. This results in the increase of the output value towards the biased classes for both ID and OOD data, therefore decreases the confidence difference D_c, *i.e.* the degradation of OOD detection performance. Motivated by WA [33] which shows the existence of biased weights in the FC classifier, we propose to perform bias correction by normalizing output logits based on the norm of weight vectors in the classifier corresponding to each learned class. Specifically, we denote the weight parameters in the classifier as $P \in \mathcal{R}^{d\times C}$ where d is the dimension of extracted feature of each input sample and C refers to the total number of classes seen so far. The weight norm of P corresponds to each learned class is calculated as

$$|W^i| = L_2(P^{1,i}, P^{2,i}, ...P^{d,i}), i \in \{1, 2, ...C\} \quad (1)$$

where $L_2()$ denotes the l_2 normalization and $P^{j,k}$ refers to the element of j^{th} row and k^{th} column in P. Let $O = \{o^1, o^2, ..., o^C\}$ denote the output from the classifier, we normalize it through

$$\hat{o}^i = o^i/|W^i|, i \in \{1, 2, ...C\} \quad (2)$$

where $\hat{o}^i$ refers to the corrected output for class i. Our weight-based normalization generates the corrected output by efficiently mitigating the bias effect from the classifier.

4.2. Confidence Enhancement

The learning objective also changes in continual learning scenario. Besides learning new tasks, we also need to maintain the learned knowledge. As shown in [21], higher confident output can decrease the model's generalization ability, which leads to catastrophic forgetting. Most existing continual learning methods address this problem by adding regularization to restrict the change of parameters [1, 3, 4, 11, 14, 23] when learning new tasks, which decrease the output confidence for both ID and OOD data, resulting in the decrease of confidence difference D_c. Our goal is to increase D_c to achieve better detection performance. Our proposed confidence enhancement method is motivated by the most recent work [6, 7], which show that the continual learning model is able to maintain the discriminativeness within each learned task. Ideally, an ID data should be more confident and task-discriminative than OOD data. Therefore, after correcting the biased output, we apply softmax on $\hat{O} = \{\hat{o}^1, \hat{o}^2, ..., \hat{o}^C\}$ to obtain $\hat{S} = \{\hat{s}^1, \hat{s}^2, ..., \hat{s}^C\}$. We extract the maximum value as $\hat{S}_{max} = max(\hat{S})$ and its corresponding task index $I_{max} = argmax_{i=1,2...K}(\hat{S})$ where K denotes the total number of tasks $\{\mathcal{T}^1, ..., \mathcal{T}^K\}$ learned so far. The softmax output value for task $\mathcal{T}_{I_{max}}$ is extracted from $\hat{S}$ as $\hat{S}_{I_{max}} = \{\hat{s}^1_{I_{max}}, \hat{s}^2_{I_{max}}, ...\hat{s}^M_{I_{max}}\}$ where M refers to the number of classes in each task, *i.e.* the incremental step size. We then measure the discriminativeness based on entropy as in Equation 3 where lower entropy H indicates more discriminative.

$$H_{I_{max}} = \sum_{i=1}^{M} \hat{s}^i_{I_{max}} \times log_M(\hat{s}^i_{I_{max}}) \quad (3)$$

Finally, we calculate the confidence score as

$$Conf = \frac{\hat{S}_{max}}{H_{I_{max}} + \epsilon} \quad (4)$$

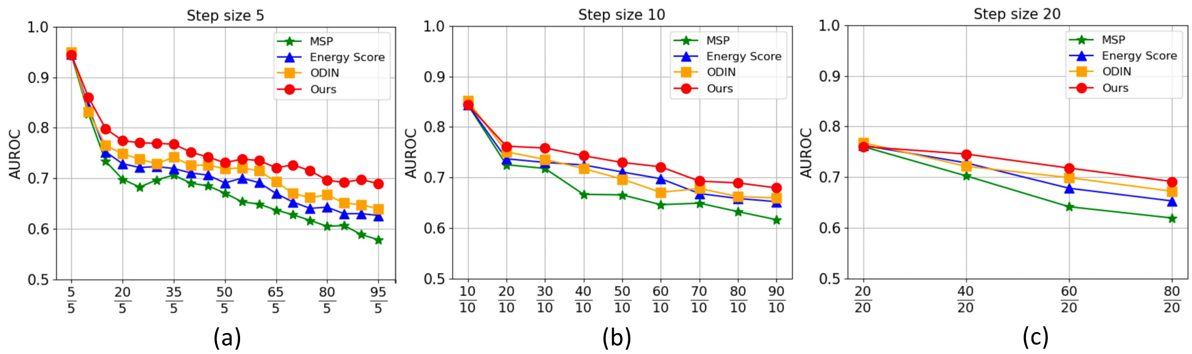

Figure 3. Results on CIFAR-100 with step size (a) 5 (b) 10 and (c) 20. The numerator and denominator of x-axis refers to the number of learned classes and new added classes, which are regarded as in-distribution and out-of-distribution data, respectively.

Methods	Step size 5			Step size 10			Step size 20		
	AUROC↑	**AUPR↑**	**FPR95↓**	**AUROC↑**	**AUPR↑**	**FPR95 ↓**	**AUROC↑**	**AUPR↑**	**FPR95↓**
MSP [10]	0.679	0.947	0.855	0.685	0.899	0.873	0.681	0.834	0.877
ODIN [15]	0.723	0.950	0.810	0.715	0.909	0.831	0.715	0.858	0.839
Energy Score [16]	0.707	0.950	0.824	0.714	0.907	0.837	0.706	0.853	0.844
Ours (w/o BC)	0.712	0.951	0.823	0.719	0.912	0.845	0.706	0.851	0.842
Ours (w/o CE)	0.708	0.947	0.836	0.713	0.907	0.851	0.699	0.844	0.854
Ours	**0.754**	**0.959**	**0.793**	**0.736**	**0.915**	**0.824**	**0.729**	**0.874**	**0.814**

Table 1. Average AUROC, AUPR and FPR95 on CIFAR-100 with step size 5, 10 and 20. BC and CE denotes bias correction step and confidence enhancement step, respectively. Best results are marked in bold.

where $\epsilon = 0.00001$ is used for regularization. Test samples assigned with larger score is regarded as ID data.

5. Experimental Results

Our proposed OOD detection method can work easily with any unsupervised continual learning approach. In this section, we show its effectiveness by incorporating the baseline in [5] to perform unsupervised continual learning. We follow the proposed evaluation protocol by comparing the OOD detection results with existing "post-hoc" methods including **MSP** [10], **ODIN** [15] and **Energy Score** [16]. We run each experiment 5 times and report the average results.

We use the CIFAR-100 [12] dataset and divide the 100 classes into splits of 20, 10 and 5 tasks with corresponding incremental step size 5, 10 and 20, respectively. For unsupervised continual learning baseline [5], we apply ResNet-32 [8] and train 120 epochs for each incremental step and the learning rate is decreased by 1/10 for every 30 epochs. Exemplar size is set as $2,000$. Following the protocol in Section 3, we perform OOD detection at the beginning of each new task except the first one.

5.1. Results on CIFAR-100

Table 1 shows the average OOD detection results on CIFAR-100 in terms of AUROC, AUPR and FPR95 as introduced in Section 3. We observe consistent improvements for OOD detection in unsupervised continual learning scenario compared with existing "post-hoc" methods. Besides, we also include **ours (w/o BC)** and **ours (w/o CE)** for ablation study where *BC* and *CE* denote bias correction and confidence enhancement steps as illustrated in Section 4. Note that the MSP [10] can be regarded as **ours (w/o BC and CE)**. Thus, both BC and CE improves the detection performance compared with MSP and our method including both steps achieve the best performance. In addition, the AUROC on CIFAR-100 for each incremental step is shown in Figure 3. Our method outperforms existing approaches at each step especially with larger margins for smaller step size, as both output bias and confidence decrease problems become more severe due to the increasing number of incremental learning steps.

6. Conclusion

In this work, we first formulate the problem of out-of-distribution detection in unsupervised continual learning (OOD-UCL) and introduce the corresponding evaluation protocol. Then a novel OOD detection method is proposed by correcting output bias and enhancing confidence difference between ID and OOD data. Our experimental results on CIFAR-100 show promising improvements compared with existing methods for various step sizes.

For future work, instead of splitting the dataset with non-overlapped classes, we will focus on unsupervised continual learning in a more realistic scenario where each new task may contain both new classes and learned classes data. Therefore, a more efficient method that can perform continual learning based on the output of OOD detection is needed for real life applications.

References

[1] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. *Proceedings of the European Conference on Computer Vision*, September 2018. 3

[2] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006. 3

[3] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13926–13935, 2020. 3

[4] Jiangpeng He and Fengqing Zhu. Online continual learning for visual food classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 2337–2346, October 2021. 3

[5] Jiangpeng He and Fengqing Zhu. Unsupervised continual learning via pseudo labels. *arXiv preprint arXiv:2104.07164*, 2021. 2, 4

[6] Jiangpeng He and Fengqing Zhu. Exemplar-free online continual learning. *arXiv preprint arXiv:2202.05491*, 2022. 3

[7] Jiangpeng He and Fengqing Zhu. Online continual learning via candidates voting. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3154–3163, January 2022. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[9] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 1

[10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 1, 2, 4

[11] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019. 3

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 4

[13] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1, 2

[14] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 1, 3

[15] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *Proceedings of International Conference on Learning Representations*, 2018. 1, 2, 4

[16] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 1, 2, 4

[17] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation. *arXiv preprint arXiv:2010.15277*, 2020. 2

[18] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989. 1

[19] Martin Mundt, Iuliia Pliushch, Sagnik Majumder, Yongwon Hong, and Visvanathan Ramesh. Unified probabilistic deep continual learning through generative replay and open set recognition. *Journal of Imaging*, 8(4):93, mar 2022. 1

[20] Jay Nandy, Wynne Hsu, and Mong Li Lee. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *Advances in Neural Information Processing Systems*, 33:9239–9250, 2020. 1

[21] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 3

[22] Dushyant Rao, Francesco Visin, Andrei A Rusu, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Continual unsupervised representation learning. *arXiv preprint arXiv:1910.14481*, 2019. 2

[23] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 2, 3

[24] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015. 3

[25] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. *International Conference on Machine Learning*, pages 8491–8501, 2020. 1, 2

[26] James Smith, Seth Baer, Cameron Taylor, and Constantine Dovrolis. Unsupervised progressive learning and the stam architecture. *arXiv preprint arXiv:1904.02021*, 2019. 2

[27] Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhanda, Ahmad Humayun, Chen Yu, Linda B Smith, and James M Rehg. Incremental object learning from contiguous views. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8777–8786, 2019. 2

[28] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. *International conference on machine learning*, pages 9690–9700, 2020. 1, 2

[29] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9302–9311, 2021. 1

[30] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019. 1, 3

[31] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1, 2

[32] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. 1

[33] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2020. 1, 3