Adapt Your Teacher: Improving Knowledge Distillation for Exemplar-free Continual Learning

Filip Szatkowski*[1,2], Mateusz Pyla[2,3,4], Marcin Przewięźlikowski[2,3,4],
Sebastian Cygert[2,5], Bartłomiej Twardowski[2,6,7], and Tomasz Trzciński[1,2,8]
[1]Warsaw University of Technology, [2]IDEAS NCBR, [3]Jagiellonian University, Faculty of Mathematics and Computer Science,
[4]Jagiellonian University, Doctoral School of Exact and Natural Sciences, [5]Gdańsk University of Technology,
[6]Autonomous University of Barcelona, [7]Computer Vision Center, [8]Tooploox

Abstract

In this work, we investigate exemplar-free class incremental learning (CIL) with knowledge distillation (KD) as a regularization strategy, aiming to prevent forgetting. KD-based methods are successfully used in CIL, but they often struggle to regularize the model without access to exemplars of the training data from previous tasks. Our analysis reveals that this issue originates from substantial representation shifts in the teacher network when dealing with out-of-distribution data. This causes large errors in the KD loss component, leading to performance degradation in CIL. Inspired by recent test-time adaptation methods, we introduce Teacher Adaptation (TA), a method that concurrently updates the teacher and the main model during incremental training. Our method seamlessly integrates with KD-based CIL approaches and allows for consistent enhancement of their performance across multiple exemplar-free CIL benchmarks.

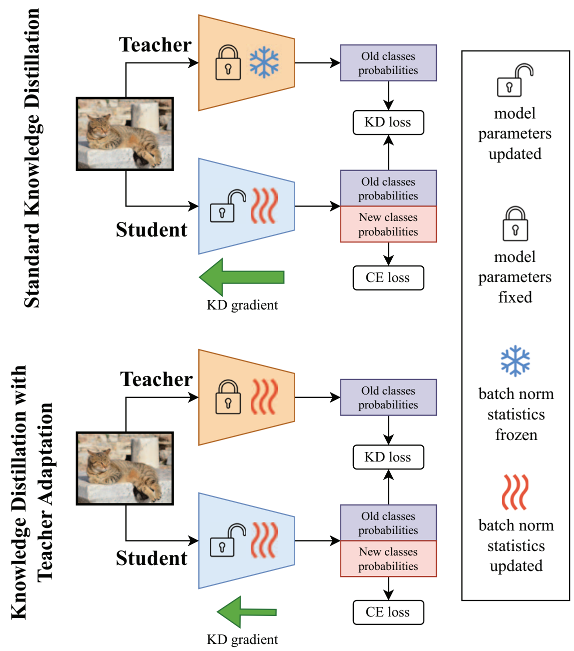

Figure 1: Comparison of vanilla Knowledge Distillation approach and our method of Teacher Adaptation. We allow the teacher model to continuously update its batch normalization statistics on the new data, which reduces knowledge distillation loss and leads to an overall more stable model.

1. Introduction

One of the most challenging continual learning scenarios is *class incremental learning* (CIL) [33, 23], where the model is trained to classify objects incrementally from the sequence of tasks, without forgetting the previously learned ones. A simple and effective method of reducing forgetting is by leveraging *exemplars* [28, 16, 5, 26] of previously encountered training examples, *e.g.* by replaying them or using them for regularization. However, this approach presents challenges, particularly in terms of additional storage needs and privacy concerns. Therefore, recently there has been a notable surge of interest in methods for more challenging exemplar-free CIL.

A common approach for exemplar-free CIL is knowledge distillation (KD), where the current model (student) is trained on the new data with a regularization term that minimizes the output difference with the previous model (teacher), which is kept frozen [21]. Since then, many methods such as iCaRL [28], EEIL [6], LUCIR [14], PodNET [11], SSIL [1], or DMC [41, 20] employed KD, but most of them use exemplars or external data.

*corresponding author, email: filip.szatkowski.dokt@pw.edu.pl

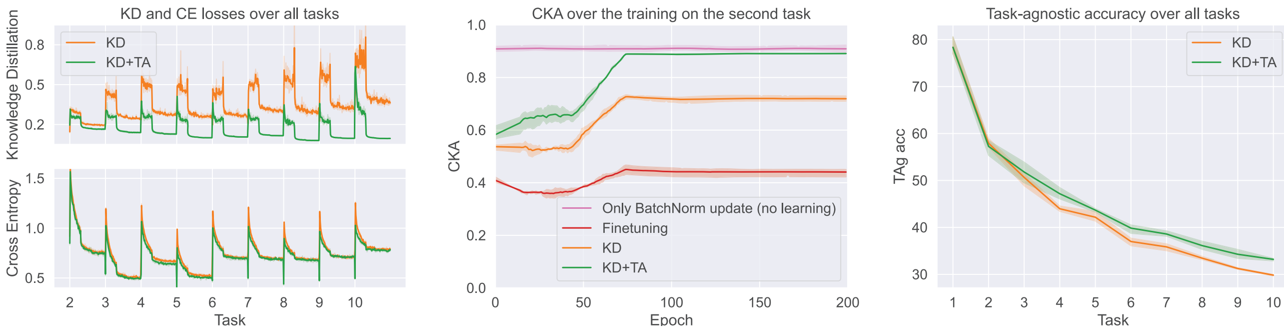

Figure 2: Applying our teacher adaptation (TA) method reduces knowledge distillation (KD) loss and improves stability over the course of continual learning. (left) KD loss and cross-entropy (CE) loss of training the model with and without TA. Our method leads to more consistent representations, as visualized by the CKA [18] between the representations of the new data obtained in the teacher and student models while learning the second task (middle). KD with TA leads to better task-agnostic accuracy (right). We conduct the experiments on CIFAR100 split into 10 tasks.

Exemplar-free CIL still remains challenging [32] for KD methods due to the possibility of significant distribution drift in subsequent tasks, which leads to large errors during training with KD loss. Motivated by the recent domain adaptation methods [34, 31], we examine the role of batch normalization (BN) statistics in CIL training with KD loss and conjecture that in standard KD methods, the KD loss between models with different BN statistics may introduce unwanted model updates due to the data distribution shifts. To avoid this, we propose to continuously adapt them to the new data for the teacher model while training the student.

We show that adapting the teacher BN statistics to the new task can significantly lower KD loss without affecting the CE loss, which leads to reduced changes in representations (Figure 2). We note that TA has been used in standard KD [43] or in the online continual learning with exemplars [12], but we are the first to apply it to exemplar-free CIL scenario, where the teacher and the model are trained on non-overlapping data.

2. Related works

Class Incremental Learning (CIL) [33, 23] aims to learn incrementally from a stream of tasks, without the knowledge about the task identifier. Most CIL methods store either exemplars or features from the previous tasks in the replay buffer [28, 16, 5, 26], modify the structure of the model [36, 35] or regularize changes in model [17, 21]. Modern CIL methods usually combine those approaches [6, 37, 29, 28, 21, 1] and often rely heavily on exemplars, which raises issues with data storage and privacy.

Regularization methods for continual learning employ either parameter regularization [17, 39, 2] or functional regularization through knowledge distillation (KD) on model activations. In CL, KD was first applied in LwF [21], and, since then, has been widely used [27, 14, 28, 26, 1, 11, 10, 40]. However, most of those methods are impractical for exemplar-free settings, as their performance heavily relies on exemplar buffer.

Modifying the teacher model in KD was recently explored in a setting where both models operate on the same domain [43, 22] and the teacher is adapted through meta-learning to better guide the student. La-MAML [12] applies a similar idea in online continual learning, using exemplars for the outer loop optimization.

Batch Normalization (BN) [15] is widely used in deep learning, but it was shown to be problematic in CL [30] as its statistics change drastically between the tasks. Alternative normalization approaches such as LayerNorm [4] or GroupNorm [38] often lead to decreased performance in standard CL models. Several domain adaptation methods use BN statistics for domain transfer [34, 31]. CL-specific normalization methods also have been proposed [25, 7], but they are not suited for exemplar-free setting.

3. Method

We propose *Teacher Adaptation* - a simple, yet effective method for CIL with KD presented in Figure 1. Our method allows the teacher model to continuously update BN statistics alongside the student when training on the new data, which addresses the problem of diverging BN statistics between the teacher and student model caused by the shifts in training data between subsequent tasks.

Knowledge Distillation in Continual Learning. Knowledge distillation (KD) methods for continual learning save the (*teacher*) model Θ_t trained after each task t and use it during learning the (*student*) model Θ_{t+1} on new task $t+1$, with general learning objective:

$$L = L_{CE} + \lambda L_{KD}, \tag{1}$$

		T10S10		T20S5		T11S50		T26S50	
		$Acc_{Inc}\uparrow$	$Forg_{Inc}\downarrow$	$Acc_{Inc}\uparrow$	$Forg_{Inc}\downarrow$	$Acc_{Inc}\uparrow$	$Forg_{Inc}\downarrow$	$Acc_{Inc}\uparrow$	$Forg_{Inc}\downarrow$
a) CIFAR100	GKD	42.52±0.76	22.26±0.31	31.89±0.45	34.68±1.87	41.69±1.18	18.09±0.88	17.64±0.93	9.67±0.26
	+*TA*	**44.09±0.97**	**19.41±0.60**	**35.99±0.79**	**23.32±1.79**	**44.05±1.12**	**12.97±0.43**	**19.37±1.73**	**8.31±0.68**
	TKD	43.74±0.84	23.65±0.79	34.58±0.34	21.13±1.17	40.44±1.40	12.20±0.46	14.64±0.33	**6.02±0.54**
	+*TA*	**45.29±1.02**	**19.42±0.85**	**34.62±0.92**	**14.72±1.28**	**41.68±1.03**	**9.29±0.75**	**16.66±1.66**	6.88±0.36
b) ImageNet100	GKD	54.62±0.52	25.95±0.11	42.82±0.58	35.39±0.88	**52.67±0.93**	**9.92±0.83**	21.91±0.06	**9.29±0.69**
	+*TA*	**55.82±0.61**	**20.52±0.24**	**45.88±0.79**	**23.25±0.62**	51.44±0.51	14.55±0.76	**22.31±0.64**	11.28±0.98
	TKD	55.70±0.49	23.55±0.35	**45.71±0.37**	25.85±0.26	**54.72±0.86**	**10.16±0.34**	19.32±0.23	**9.67±0.61**
	+*TA*	**56.23±0.70**	**18.09±0.26**	45.14±0.78	**15.62±0.51**	53.85±0.39	13.15±0.16	**22.55±0.83**	9.96±0.28

Table 1: Comparision of standard Knowledge Distillation (KD) techniques with added Teacher Adaptation (TA) on different splits of a) CIFAR100 and b) ImageNet100. Adapting the teacher is beneficial to the learning process for all the tasks.

where L_{CE} is the cross-entropy loss, L_{KD} is the KD loss and λ is the coefficient that controls the trade-off between stability and plasticity.

The most popular formulation of KD loss was proposed in [21]. Following [1], we refer to it as global KD (GKD) and define it as:

$$\mathcal{L}_{GKD}(\mathbf{y}_o, \hat{\mathbf{y}}_o) = -\sum_{i=1}^{|C_t|} p_o^{(i)} \log \hat{p}_o^{(i)}, \tag{2}$$

where $|C_t|$ is the number of classes learned by previous model Θ_t and $p_o^{(i)}$, $\hat{p}_o^{(i)}$ are temperature-scaled softmax probabilities:

$$p_o^{(i)} = \frac{e^{y_o/T}}{\sum_j e^{y_o/T}}, \quad \hat{p_o}^{(i)} = \frac{e^{\hat{y_o}/T}}{\sum_j e^{\hat{y_o}/T}} \tag{3}$$

We denote temperature parameter with T and use o to emphasise that the logits $y_o^{(i)}$, $\hat{y}_o^{(i)}$ only relate to *old* classes from previous tasks.

Ahn et al. [1] noticed that GKD formulation encourages forgetting of previous tasks and proposed task-wise knowledge distillation (TKD), which computes softmax probabilities separately across the model classification heads:

$$\mathcal{L}_{TKD}(\mathbf{y}_o, \hat{\mathbf{y}}_o) = \sum_{i=1}^{t} \mathcal{D}_{KL}(p_o^{(i)} \log \hat{p}_o^{(i)}), \tag{4}$$

where $\mathcal{D}_{KL}$ is Kullback–Leibler divergence and $p_o^{(i)}$, $\hat{p}_o^{(i)}$ are computed task-wise across the outputs for task i as in Equation (3)).

Teacher Adaptation. Most models used in CIL for vision tasks are convolutional neural networks such as ResNet [13], which typically use BN layers and keep the parameters and statistics of those layers in the teacher model Θ_t fixed during learning Θ_{t+1}. However, when changing the task, BN statistics in both models quickly diverge, which leads to higher KD loss. Gradient updates in this case not only regularize the model towards the previous tasks, but also compensate for the changes in BN statistics, harming the learning process.

Inspired by test-time adaptation methods [34], we propose to reduce this negative interference with a simple method that we label *Teacher Adaptation (TA)*. Our method updates BN statistics of both models simultaneously on new data while learning the new task. As shown in Figure 2, it allows for significantly reduced KD loss over learning from sequential tasks in CIL, which improves the overall model stability.

4. Experiments

TA on standard CIL benchmarks. We evaluate knowledge distillation approaches described in Section 3 on the standard continual learning benchmarks CIFAR100 [19] and ImageNet-Subset [9], each containing images from 100 classes. For experiments on CIFAR100, we keep the class order from iCaRL [28] and we use ResNet32 [13]. For ImageNet Subset, we use ResNet18 [13]. We investigate different class splits, which we denote using the total number of tasks **T** (which includes the first pretraining task if present), and the number of classes in the first task **S**. We use FACIL framework provided by Masana et al. [23], and always use the same hyperparameters for each KD method within a single setting. We train the network on each new task for 200 epochs in all experiments, using SGD optimizer without momentum or weight decay. Following Zhou et al. [42], we use a learning rate scheduler with the initial learning rate of 0.1 and 10x decay on the 60th, 120th, and 160th epoch. We report the results averaged over three random seeds. We provide the description of reported metrics in Appendix.

We present the results obtained on standard CIL baselines in Table 1. On CIFAR100, TA consistently improves the accuracy across all the settings. On ImageNet, our method improves upon the baseline for most settings, or at worst stays within the margin of error of the baseline. We observe that applying our method generally leads to a more stable network and therefore reduces forgetting, *i.e.* TKD+TA for equally split ImageNet (T10S10, T20S5).

Teacher Adaptation under varying degrees of distribution shift. We also introduce a corrupted CIFAR100 setting where data in every other task contains a noise of varying

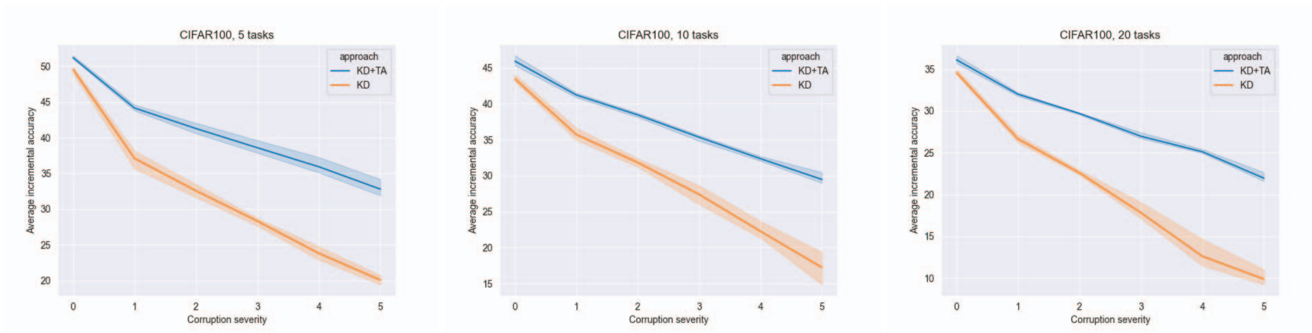

Figure 3: Average incremental accuracy for standard KD and our method of TA under varying strength of data shift on splits of CIFAR100. We vary the shift strength by adding Gaussian noise of different severity to every other task. As the noise strengthens, the gap between TA and standard KD widens. Our method leads to more robust learning in case of data shifts.

severity, which allows us to measure the impact of TA under varying and controllable degrees of data shift. We corrupt every other task in this setting with Gaussian noise, so that in subsequent tasks the data distribution changes from clean to noisy or vice versa. We obtain varying strength of distribution shift by using different levels of noise severity, following the methodology from [24]. We show the results of this experiment in Figure 3. As the noise severity increases, the gap between standard KD and TA widens, indicating that our method is better suited to more challenging scenarios of learning under extreme data distribution shifts.

Alternative solutions to problems with batch normalization. To justify the validity of our method, we compare it with other potential solutions for the problem with BN layers. We use GKD on CIFAR100 split into 10 tasks and compare the following solutions: 1) standard training with BN statistics from the previous task fixed in the teacher model, but updated in the student model, 2) BN layers removed, 3) BN statistics fixed in both models after learning the first task, 4) BN layers replaced with LayerNorm [4] layers, and 5) finally our solution of Teacher Adaptation. We show the results of those experiments in Table 2. Fixing or removing BN leads to unstable training, which can be partially fixed by setting a high gradient clipping value or lowering the lambda parameter, but at the cost of the worse performance of the network. Training the networks with LayerNorm is stable, but ultimately those networks converge to much worse solutions than the variants with BN. Our solution is the only one that improves over different values of λ and without the need of clipping the gradient values.

$clip = 100$	$\lambda = 5$		$\lambda = 10$	
	Acc_{Final}	Acc_{Inc}	Acc_{Final}	Acc_{Inc}
1) **GKD**	25.47±0.57	41.59±0.32	27.96±0.34	42.28±0.67
2) **-BN**	0.33±1.15	2.01±2.67	0.33±1.15	2.85±3.81
3) **fix BN**	-	-	-	-
4) **-BN +LN**	21.94±0.95	34.7±0.48	22.76±1.05	34.48±0.15
5) **+TA**	**31.39±0.17**	**44.98±0.38**	**31.85±0.10**	**44.06±0.69**
$clip = 1$	$\lambda = 5$		$\lambda = 10$	
	Acc_{Final}	Acc_{Inc}	Acc_{Final}	Acc_{Inc}
1) **GKD**	20.80±0.56	34.28±0.24	27.96±0.34	42.28±0.67
2) **-BN**	19.47±0.18	29.83±0.53	0.33±1.15	2.85±3.81
3) **fix BN**	20.21±0.31	32.07±0.20	-	-
4) **-BN +LN**	18.49±1.41	30.39±0.72	22.76±1.05	34.48±0.15
5) **+TA**	**24.19±0.90**	**36.13±0.24**	**31.85±0.10**	**44.06±0.69**

Table 2: Results for different solutions to the problem of diverging BN layers when using KD in CL. "-" indicates that training crashes due to instability. TA is the only solution that improves upon the baseline.

5. Conclusions

We propose Teacher Adaptation (TA), a simple, yet effective method to improve the performance of KD-based methods in exemplar-free CIL. During learning a new task, TA updates the teacher network by adjusting its BN statistics with new data. This mitigates the changes in the model caused by KD loss that arise as the current learner constantly tries to compensate for the diverging normalization statistics between itself and the teacher model. We show that TA consistently improves the results for different KD-based methods on several CIL benchmarks in an exemplar-free setting. Moreover, we demonstrate that benefits from our method increase as we increase the degree of shift in data between subsequent tasks. TA can be easily added to the existing CIL methods and induces only a slight computational overhead, making it a valuable addition to existing exemplar-free KD-based CIL methods.

Acknowledgements

Filip Szatkowski and Tomasz Trzcinski are supported by National Centre of Science (NCP, Poland) Grant No. 2022/45/B/ST6/02817. Tomasz Trzciński is also supported by NCP Grant No. 2020/39/B/ST6/01511.

References

[1] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 844–853, 2021.

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget, 2018.

[3] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *CVPR*, pages 3366–3375, 2017.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[5] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8218–8227, 2021.

[6] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018.

[7] Sungmin Cha, Sungjun Cho, Dasol Hwang, Sunwon Hong, Moontae Lee, and Taesup Moon. Rebalancing batch normalization for exemplar-based class-incremental learning, 2023.

[8] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, pages 532–547, 2018.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009.

[10] Fei Ding, Yin Yang, Hongxin Hu, Venkat Krovi, and Feng Luo. Multi-level knowledge distillation via knowledge alignment and correlation, 2021.

[11] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 86–102. Springer, 2020.

[12] Gunshi Gupta, Karmesh Yadav, and Liam Paull. Look-ahead meta learning for continual learning. *Advances in Neural Information Processing Systems*, 33:11588–11598, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[16] Ahmet Iscen, Thomas Bird, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. A memory transformer network for incremental learning, 2022.

[17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[18] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[20] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 312–321, 2019.

[21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[22] Xinge Ma, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. Knowledge distillation with reptile meta-learning for pretrained language model compression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4907–4917, 2022.

[23] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[24] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[25] Quang Pham, Chenghao Liu, and Steven Hoi. Continual normalization: Rethinking batch normalization for online continual learning. *arXiv preprint arXiv:2203.16102*, 2022.

[26] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 524–540. Springer, 2020.

[27] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018.

[28] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[29] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[30] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization?, 2019.

[31] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551. Curran Associates, Inc., 2020.

[32] James Seale Smith, Junjiao Tian, Shaunak Halbe, Yen-Chang Hsu, and Zsolt Kira. A closer look at rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2409–2419, 2023.

[33] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[34] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[35] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 631–648. Springer, 2022.

[36] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[37] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] Yuxin Wu and Kaiming He. Group normalization, 2018.

[39] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence, 2017.

[40] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C. C. Jay Kuo. Class-incremental learning via deep model consolidation, 2020.

[41] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry P. Heck, Heming Zhang, and C.-C. Jay Kuo. Class-incremental learning via deep model consolidation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1120–1129. IEEE, 2020.

[42] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, and De-Chuan Zhan. Pycil: A python toolbox for class-incremental learning, 2023.

[43] Wangchunshu Zhou, Canwen Xu, and Julian J. McAuley. BERT learns to teach: Knowledge distillation with meta learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7037–7049. Association for Computational Linguistics, 2022.