

Resurrecting Old Classes with New Data for Exemplar-Free Continual Learning

Dipam Goswami^{1,2} Albin Soutif-Cormerais^{1,2} Yuyang Liu^{3,4,†} Sandesh Kamath^{1,2}
Bartłomiej Twardowski^{1,2,5} Joost van de Weijer^{1,2}

¹Department of Computer Science, Universitat Autònoma de Barcelona

²Computer Vision Center, Barcelona ³University of Chinese Academy of Sciences

⁴Shenyang Institute of Automation, Chinese Academy of Sciences ⁵IDEAS-NCBR

{dgoswami, albin, skamath, btwardowski, joost}@cvc.uab.es, liuyuyang@sia.cn

Abstract

Continual learning methods are known to suffer from catastrophic forgetting, a phenomenon that is particularly hard to counter for methods that do not store exemplars of previous tasks. Therefore, to reduce potential drift in the feature extractor, existing exemplar-free methods are typically evaluated in settings where the first task is significantly larger than subsequent tasks. Their performance drops drastically in more challenging settings starting with a smaller first task. To address this problem of feature drift estimation for exemplar-free methods, we propose to adversarially perturb the current samples such that their embeddings are close to the old class prototypes in the old model embedding space. We then estimate the drift in the embedding space from the old to the new model using the perturbed images and compensate the prototypes accordingly. We exploit the fact that adversarial samples are transferable from the old to the new feature space in a continual learning setting. The generation of these images is simple and computationally cheap. We demonstrate in our experiments that the proposed approach better tracks the movement of prototypes in embedding space and outperforms existing methods on several standard continual learning benchmarks as well as on fine-grained datasets. Code is available at <https://github.com/dipamgoswami/ADC>.

1. Introduction

Deep learning has gained widespread use in various computer vision tasks, demonstrating exceptional performance when trained on a dataset in a single session. However, a significant challenge arises when new data is introduced incrementally in multiple phases or tasks. Then neural networks need to adapt without forgetting previously

[†] Yuyang Liu is the corresponding author.

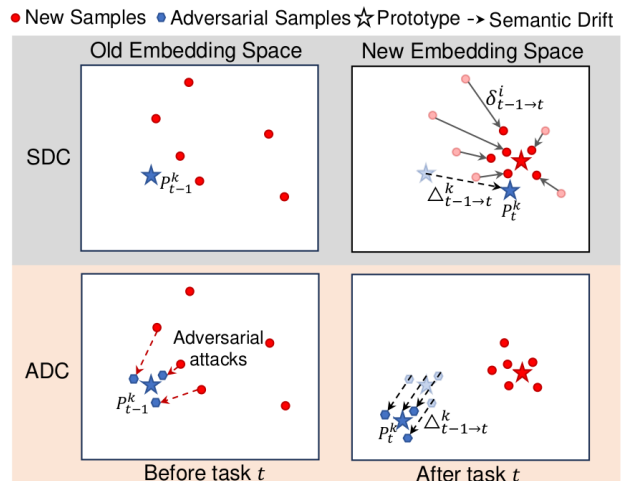


Figure 1. Illustration of Adversarial Drift Compensation (ADC) and SDC [52]. In SDC, the drift $\Delta_{t-1 \rightarrow t}^k$ is estimated as the average of drift of all new task samples after training on a new task. Instead, we propose to move the new task features close to the old prototype P_{t-1}^k of class k by perturbing the new images using targeted adversarial attacks. The drift of the adversarial samples from old to new feature space is used to resurrect all old prototypes.

learned information, a phenomenon known as *catastrophic forgetting* [18, 30]. Recent studies in continual learning (CL) [4, 29, 50, 54] focus on two prevalent scenarios [47]: Task-Incremental Learning (TIL), where task information is available during testing, and Class-Incremental Learning (CIL), where it is not. Our work aims to address the more challenging CIL problem.

Exemplar-based CIL methods [2, 3, 6, 8, 13, 25, 38, 49] store small subsets of data from each task. These exemplars are later replayed with current data during training in new tasks. Although effective, these methods necessitate storing input data from previous tasks, leading to multiple challenges in practical settings such as legal concerns with new

regulations (e.g. European GDPR where users can request to delete personal data), and privacy issues when dealing with sensitive data like in medical imaging. Recently, the *exemplar-free CIL* (EFCIL) setting is extensively studied [11, 28, 35, 52, 55–57]. However, unlike exemplar-based methods, the EFCIL methods are only effective when starting with high-quality feature representations and are thus dependent on having a large initial task which is typically half of the whole dataset. However, a more practical CIL approach should be able to perform well on training from a smaller initial task and at the same time should not store exemplars. We define this as a *small-start setting* and analyze how existing EFCIL methods perform in this setting.

A critical aspect in CIL is the *semantic drift* of feature representations [52] after training on new tasks. This results in the movement of class distributions in feature space. Thus, it is crucial to track the old class representations after learning new tasks. While the class-mean in the new feature space can be effectively estimated using Nearest-Mean of Exemplars (NME) [8, 36], it is challenging to estimate it without exemplars. Usually, this drift is minimized with heavy functional regularization, which consequently restricts the plasticity of the network. Another way is to estimate it from the drift of current data, as done in SDC [52] or by augmenting old prototypes using new class features [28, 42]. In this paper, we propose a novel drift estimation method using adversarial examples to resurrect old class prototypes in the new feature space as shown in Fig. 1.

Adversarial examples [1, 14, 27, 31, 45] are maliciously crafted inputs that are designed to fool a neural network into predicting a different output than the one initially predicted for the original input. Exploiting the concept of targeted adversarial attacks [22, 27], we propose to perturb the new data such that the adversarial images result in embeddings close to the old prototypes. Now, the drift from old to new feature space is estimated using these adversarial samples, which serve as pseudo-exemplars for the old classes. We hypothesize that the pseudo-exemplars behave like the original exemplars in the feature space, and thus we exploit them to measure the drift. This generation of adversarial samples is computationally cheaper and much faster (only a few iterations) compared to data-inversion methods [51] which inverts embeddings to realistic images.

Following recent studies [11, 16, 26], we explore using class prototypes with an NCM [36] classifier and show that a simple baseline of logits distillation [24] with an NCM classifier often outperforms existing EFCIL methods in the small-start setting. Applying our proposed drift compensation method with this baseline, we obtain state-of-the-art performance with significant gains over existing methods on standard CL benchmarks using CIFAR-100 [20], Tiny-ImageNet [23] and ImageNet-Subset [5] as well as fine-grained datasets like CUB-200 [48] and Stanford Cars [19].

Our contributions can be summarized as:

- We study the challenging EFCIL settings and highlight the importance of continually learning from small-start settings instead of assuming the availability of half of the dataset in the first task.
- We present a novel and intuitive method - Adversarial Drift Compensation (ADC) to estimate semantic drift and resurrect old class prototypes in the new feature space. We also investigate how adversarially generated samples transfer in CIL settings from old to the new model.
- We perform experiments on several CIL benchmarks and outperform state-of-the-art methods by a large margin on several benchmark datasets. Especially notable are our results on fine-grained datasets, where we report performance gains of around 9% for last task accuracy.

2. Related Work

Class-Incremental Learning. CIL [4, 29, 54] methods aim to learn new data which arrives incrementally and suffers from the catastrophic forgetting problem [30, 37]. During evaluation in CIL without the task id, it is difficult to distinguish classes that belong to different tasks [44]. While in general this setting is tackled using *rehearsal approaches* [2, 6, 8, 13, 38] by storing raw inputs, some attempts have been made without storing raw inputs. LwF [24] prevents important changes in the network by preventing the output of the current model to drift too much from the output of the previous model. PASS [55] learns the backbone using self-supervised learning and later uses functional regularization and feature rehearsal, SSRE [56] proposed an architecture organization strategy that aims to transfer invariant knowledge across tasks. In FeTRIL [35], the authors freeze the feature extractor and estimate the position of the old class features by using the current task data variance. Recently, FeCAM [11] leveraged the mean and covariance of the previous task features and proposed a Mahalanobis distance-based classifier.

Drift estimation. When updating the feature extractor on new classes, the representation learned for the old class prototypes changes and thus the need to rectify those drifts [52]. SDC [52] showed that the new data can be used to estimate the drift of the old prototype representations. Recent methods [28, 42, 46] also explored how to update the prototypes learned in old tasks to counter the drift. Toldo et al. proposed to learn the relations between old and new class features to estimate the drift. NAPA-VQ [28] proposed to augment the prototypes using the topological information of classes in the feature space. Prototype Reminiscence [42] proposed to dynamically reshape old class feature distributions by interpolating the old prototypes with the new sample features. In this work, we generate adversarial samples which behaves as pseudo-exemplars and is then used to measure the drift.

CL using Adversarial Attacks. Adversarial Attacks has been studied in-depth in recent years [1, 14, 45], and has been later harnessed to create realistic looking images from a trained vision model [33], including inputs that can be later used for training [51]. Some recent methods [9, 17, 21, 43] in exemplar-based CIL borrowed the idea of adversarial attacks. ASER [43] used the kNN-specific Shapley value to obtain more representative buffer samples. GMED [17] edits the exemplars by monitoring the change in loss when training on incoming data. RAR [21] used the pairwise relations between the exemplars and the new samples and perturb the exemplars to obtain samples close to the decision boundaries. While all these approaches use adversarial attacks on the memory samples, we use it to perturb the new data to simulate the old data.

3. Method

We consider the EFCIL setup where new classes emerge over time and we are not allowed to store samples from old classes. These classes come in different tasks, one task at a time, and the tasks contain a mutually exclusive set of classes. When training on task t , we have access to current dataset $D_t = \{X_t, Y_t\}$ with images X_t and labels Y_t . The main goal of EFCIL is to learn a model h that correctly classifies the data into classes encountered so far. We use $h_t(x) = \sigma(W_t f_t(x))$, where f_t is the feature extractor parameterized by θ_t learned in task t and W_t is weight matrix of the linear classifier with softmax function σ .

3.1. Motivation

In general, for a new feature extractor f_t trained on new data and an old feature extractor f_{t-1} from previous task, we have access to old class prototypes up to task $t - 1$ denoted by $P_{t-1}^{Y_{1:t-1}}$. We compute the prototypes for all new classes after training in the current task. For a class k in task t , we compute $P_t^k = \frac{1}{|X_t^k|} \sum_{x \in X_t^k} f_t(x)$, where X_t^k is the set of samples from class k , $f_t(x)$ is the feature embedding for an image x from class k . However, the old class prototypes were computed on the old feature space $f_{t'}$ ($t' < t$) in old tasks and have drifted to a different position in the new feature space f_t after training on new data.

Previously, SDC [52] proposed to compensate the drift of these prototypes $P_{t-1}^{Y_{1:t-1}}$ by computing the drift from old model embeddings $f_{t-1}(x)$ to new model embeddings $f_t(x)$ corresponding to all images x in the current task. This drift of the current data is then used to approximate the drift of previous task prototypes by considering a weighted Gaussian window around the prototypes (giving more weights to drift vectors close to the prototype). However, the quality of this drift approximation for previous prototypes is expected to be low when few current task data points are close to a given prototype in the embedding

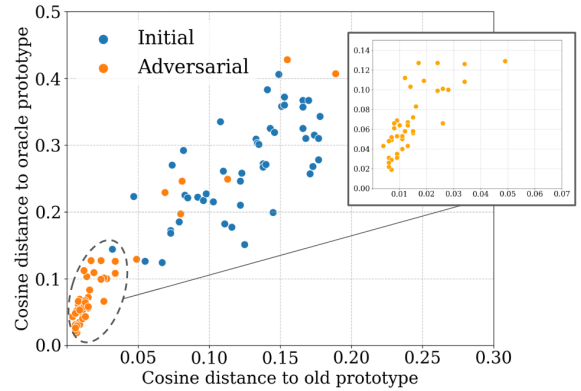


Figure 2. Illustration to show that the cosine distance between embeddings and old prototype in the old feature space is correlated with the cosine distance between embeddings and oracle prototype in the new feature space. This holds true for embeddings of both initial and adversarial samples. For demonstration, we select few current-task samples that are closest to the old prototype and choose the same target old class for all samples. The blue and orange points represents the non-modified current class samples and the modified samples using our proposed approach respectively. In this analysis, we compute the oracle prototype using all old task data in the new feature space.

space. We show in our experiments that SDC indeed struggles to estimate the drift in the small-start settings where the feature representations change considerably.

In a similar fashion, it is possible to estimate this drift without using such weights, by simply choosing the closest samples to each old class prototype and compute the average feature of these samples when fed to the new backbone. We select samples from the current task that are close to the old prototype of a given class and verify that such samples also lie close to the oracle prototype (in the new feature space). We analyze in Fig. 2 that there exist a correlation between the distance to the old prototype in the old feature space and the distance to the oracle prototype in the new feature space (see blue dots). This motivates us to leverage current task samples so that their distance to the old prototype in the old feature space is even smaller, which could in turn improve the drift estimation. We hypothesize that this can be done by computing adversarial samples from the current task samples, aiming for their representation to match one of the old class prototypes in the old feature space.

3.2. Adversarial Drift Estimation

To estimate the drift of old class prototypes after updating the model on new classes, it is desirable to have the exemplars. These exemplars can be passed through the new model to compute the *oracle* prototype position in the new feature space. However, in the exemplar-free setting, we can only access the new data. In order to use the new data to represent the old data, we exploit the concept of targeted

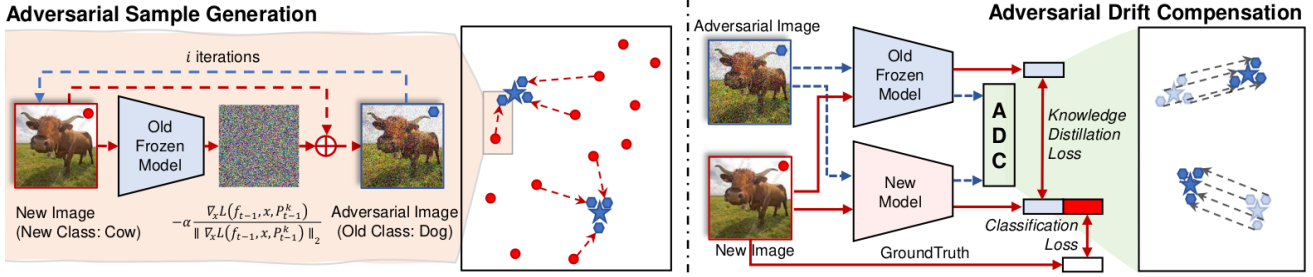


Figure 3. (a) Adversarial Sample Generation: On the old model feature space, the new samples closest to the old prototype are selected and iteratively perturbed in the direction of the target old prototype to generate adversarial samples which are now misclassified as the target old class resulting in embeddings closer to the old prototype. We perform this for every old class (we show 2 classes here for demonstration). (b) Model Training with Drift Compensation: The new model is trained using the classification loss for learning new classes and knowledge distillation loss to prevent forgetting of old classes. After the new model is trained, the adversarial samples generated using the old model are passed through both the models and the drift from old to new feature space is estimated. This is then used to update the old prototypes.

adversarial attacks [22, 27] to target one old class at a time and perturb the new data in a way that it serves as a substitute of old data to the model. We perform adversarial attacks on new data to move its embeddings very close to old prototypes in the old feature space. Using the adversarial samples, we can estimate the drift from old to new feature space and compensate it as illustrated in Fig. 3.

To estimate the drift of prototype P_{t-1}^k for a target old class k , we obtain \mathcal{X}^k by sampling a set of m data points from the current task data X_t which are closest to P_{t-1}^k based on L2 distance between the embeddings of samples in X_t and the prototype P_{t-1}^k . We aim to perturb the samples $x \in \mathcal{X}^k$ and obtain \mathcal{X}_{adv}^k such that the adversarial samples $x_{adv} \in \mathcal{X}_{adv}^k$ are closer to P_{t-1}^k and are now classified to class k using the NCM classifier in the old feature space:

$$k = \underset{y \in Y_{1:t-1}}{\operatorname{argmin}} \|f_{t-1}(x_{adv}) - P_{t-1}^y\|_2. \quad (1)$$

We propose the following optimization objective by computing the mean squared error between the features $f_{t-1}(x)$ and the prototype P_{t-1}^k as:

$$L(f_{t-1}, \mathcal{X}^k, P_{t-1}^k) = \frac{1}{|\mathcal{X}^k|} \sum_{x \in \mathcal{X}^k} \|f_{t-1}(x) - P_{t-1}^k\|_2^2. \quad (2)$$

In order to move the feature embedding in the direction of the target prototype P_{t-1}^k , we obtain the gradient of the loss with respect to the data $x \in \mathcal{X}^k$, normalize it to get the unit attack vector and scale it by α as follows:

$$x_{adv} \leftarrow x - \alpha \frac{\nabla_x L(f_{t-1}, x, P_{t-1}^k)}{\|\nabla_x L(f_{t-1}, x, P_{t-1}^k)\|_2} \quad \forall x \in \mathcal{X}^k \quad (3)$$

where $\nabla_x L(f_{t-1}, x, P_{t-1}^k)$ is the gradient of the objective function with respect to the data x and α refers to the step size. We perform the attack for i iterations.

Here, the goal is different from conventional adversarial attacks like FGSM and its variants [7, 10, 22, 27] which aim to minimize the perturbation in order to keep the perturbed image visually similar to the real image by having a fixed ϵ -budget generally based on ℓ_2 or ℓ_∞ -norm of perturbation. In our case, we do not need to apply such restrictions on the distance between initial and final image, instead, we only clip the perturbed image in the existing range of pixel values. We show in supplementary materials that indeed the generated adversarial images have much higher perturbation. We do observe that our formulation is closer to the ℓ_2 -norm based attack as we use ℓ_2 normalization of the gradient vector to obtain a unit perturbation vector which is scaled using the step size.

Continual Adversarial Transferability. An interesting aspect of our method is that the adversarial samples x_{adv} are crafted on the old feature extractor f_{t-1} and then passed to the new feature extractor f_t expecting that the adversarial samples will still be misclassified as the target old class k . We define this as *continual adversarial transferability* where the adversarial samples generated on the old feature space still behave in the same way on the new feature space. This is feasible since the old model and the new model are not entirely different because of the knowledge distillation used in order to reduce catastrophic forgetting. This is related with the concept of adversarial transferability [15, 27, 32, 34] where an attack obtained on one neural network also behaves as an attack on other independently trained neural network based architectures.

We analyze the oracle setting using the old class data to validate the continual adversarial transferability. We show in Fig. 2 that distance of the adversarial samples from their target prototypes in the old feature space is still correlated with their distance to the oracle prototypes of the target class in the new feature space. This suggests that the adversarial samples crafted using the old feature space are still effective in the new feature space and therefore allows us to reliably

compute the drift from these adversarial samples.

3.3. Drift Compensation

The adversarial samples when passed through the new feature extractor f_t are expected to lie close to the drifted prototype and hence are used to compute the drift. After generating the adversarial samples for each target class k , we measure the prototype drift as:

$$\Delta_{t-1 \rightarrow t}^k = \frac{1}{|\mathcal{X}_{adv}^k|} \sum_{x_{adv} \in \mathcal{X}_{adv}^k} (f_t(x_{adv}) - f_{t-1}(x_{adv})) \quad (4)$$

where $x_{adv} \in \mathcal{X}_{adv}^k$ is the set of only those adversarial samples which are classified as the target class k using the NCM classifier. We resurrect the old prototypes by compensating the drift as follows:

$$P_t^k = P_{t-1}^k + \Delta_{t-1 \rightarrow t}^k \quad (5)$$

After compensating all old prototypes, we use the NCM classifier in the new feature space for classifying the test samples. Unlike SDC [52], we do not perform weighted averaging based on the distances to the prototype since embeddings from adversarial images are very close to the prototypes and we found no gain by applying this additional weighting scheme.

3.4. Training Strategy

In addition to learning new classes, we perform knowledge distillation [24] on the logits to transfer knowledge from the frozen teacher model at previous task $t - 1$ to the student model at current task t as follows:

$$\mathcal{L} = \mathcal{L}_{ce}(h_t(x), y) + \lambda \mathcal{L}_{ce}(h_{t-1}(x), h_t(x)) \quad (6)$$

where λ refers to the regularization strength, the first term \mathcal{L}_{ce} refers to the cross-entropy loss for learning new classes and the second term performs the regularization by forcing the probabilities of old classes on the old model h_{t-1} and new model h_t to be similar and thus prevents forgetting.

4. Experiments

Datasets. We perform experiments on several CIL benchmarks. CIFAR-100 [20] contains 50k training images of size 32x32 and 10k test images, divided in 100 classes. TinyImageNet[23] contains 100k training images and 10k test images from 200 classes and image size of 64x64, taken as a subset of ImageNet [5]. ImageNet-Subset is a subset of the ImageNet (ILSVRC 2012) dataset [39] containing 100 classes with a total of 130k training images and 5k test images and image size of 224x224. We equally split all these datasets in 5 and 10 tasks. This is different from the

Algorithm 1 Adversarial Drift Compensation

input: Images X_t , feature extractors f_t and f_{t-1} , old prototypes $P_{t-1}^{Y_{1:t-1}}$, step size α , number of generated samples m , number of iterations for perturbation i .
output: Resurrected prototypes $P_t^{Y_{1:t-1}}$.

- 1: **for** $k \in Y_{1:t-1}$ **do**
- 2: Sample a set \mathcal{X}^k of m new samples from X_t which are closest to P_{t-1}^k based on L2 distance.
- 3: **for** i iterations **do**
- 4: $L(f_{t-1}, \mathcal{X}^k, P_{t-1}^k) = \frac{1}{|\mathcal{X}^k|} \sum_{x \in \mathcal{X}^k} \|f_{t-1}(x) - P_{t-1}^k\|_2^2$.
- 5: $x_{adv} = x - \alpha \frac{\nabla_x L(f_{t-1}, x, P_{t-1}^k)}{\|\nabla_x L(f_{t-1}, x, P_{t-1}^k)\|_2} \forall x \in \mathcal{X}^k$
- 6: $x \leftarrow x_{adv}$
- 7: **end for**
- 8: add x in \mathcal{X}_{adv}^k if $\operatorname{argmin}_{y \in Y_{1:t-1}} \|f_{t-1}(x) - P_{t-1}^y\|_2 = k$.
 ▷ Store only those samples in \mathcal{X}_{adv}^k which are successfully misclassified as k
- 9: $\Delta_{t-1 \rightarrow t}^k = \frac{1}{|\mathcal{X}_{adv}^k|} \sum_{x_{adv} \in \mathcal{X}_{adv}^k} f_t(x_{adv}) - f_{t-1}(x_{adv})$
- 10: $P_t^k = P_{t-1}^k + \Delta_{t-1 \rightarrow t}^k$ ▷ Prototype resurrection
- 11: **end for**

big-start settings with half of the dataset in first task, commonly used in EFCIL benchmarks [11, 35, 55]. We also use two fine-grained datasets for our experiments. CUB-200 [48] contains 200 classes of birds with 224x224 image size, 5994 images for training and 5794 images for testing. We use the 5-split and 10-split settings for CUB-200. Stanford Cars [19] consists of 196 car models with 224x224 images, 8144 for training and 8041 for testing and we split it into 7 and 14 tasks.

Training Details. We use the PyCIL framework [53] as a basis for all our experiments. The training is performed using the ResNet18 model [12] and the SGD optimizer. For CIFAR-100, in the first task, we use a starting learning rate of 0.1, momentum of 0.9, batch size of 128 and weight decay of 5e-4 for 200 epochs, the learning rate is reduced by a factor of 10 after 60, 120, and 160 epochs. In the subsequent tasks, we use an initial learning rate of 0.05 reduced by a factor of 10 after 45 and 90 epochs and train for 100 epochs. Following [29], we set the regularization strength to 10 and the temperature to 2. The network is trained from scratch on CIFAR-100, TinyImageNet and ImageNet-Subset. For the experiments on fine-grained datasets, we use the ImageNet pretrained weights following standard practice [40, 52]. For ADC, we use a α value of 25, iterations $i = 3$ and number of closest samples $m = 100$ for all the datasets. Similar to most existing methods, we store all the class prototypes. Complete details about the training setting for all the datasets are given in the supplementary materials.

Method	CIFAR-100				TinyImageNet				ImageNet-Subset			
	T = 5		T=10		T = 5		T =10		T = 5		T = 10	
	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}
LwF [24]	45.35	61.94	26.14	46.14	38.81	49.70	<u>27.42</u>	38.77	50.88	69.11	37.90	61.60
NCM	53.53	<u>66.35</u>	41.31	57.85	38.69	50.45	<u>26.56</u>	<u>41.04</u>	57.74	71.99	<u>45.86</u>	65.04
SDC [52]	<u>54.94</u>	64.82	<u>41.36</u>	<u>58.02</u>	<u>40.05</u>	<u>50.82</u>	27.15	40.46	<u>59.82</u>	<u>74.10</u>	43.72	<u>65.83</u>
PASS [55]	49.75	63.39	37.78	52.18	36.44	48.64	26.58	38.65	50.96	66.15	38.90	54.74
SSRE [56]	42.39	56.57	29.44	44.38	30.13	43.20	22.48	34.93	40.30	57.57	28.12	45.87
FeTrIL [35]	45.11	60.42	36.69	52.11	29.91	43.99	23.88	36.35	49.18	63.83	40.26	55.12
FeCAM [11]	47.28	61.37	33.82	48.58	25.62	39.85	23.21	35.32	54.18	67.21	42.68	57.45
ADC (Ours)	59.14	69.62	46.48	61.35	41.0	50.94	32.32	43.04	62.40	74.84	47.58	67.07

Table 1. Evaluation of EFCIL methods on small-start settings. Best results in **bold** and second best results are underlined.

Compared Methods. Since none of the EFCIL methods are designed to start from a small first task, we implement those methods in our small-start settings. This includes LwF [24], PASS [55], SSRE [56], FeTrIL [35] and FeCAM [11]. Naturally, we also include a comparison to the existing drift-estimation method SDC [52] and the baseline model with NCM classifier. For SDC and NCM results reported in Tab. 1 and Tab. 2, we train the models using LwF and perform NCM classification in the feature space. For FeTrIL and FeCAM, the feature extractor is frozen after the first task, while for the other methods, it is continually learned. Note that here we adapt SDC with distillation on the logits, which is different from [52] where they performed distillation on the features.

Evaluation. We report the average accuracy after the last task denoted by A_{last} and the average incremental accuracy which is the average of the accuracy after all tasks (including the first one) denoted by A_{inc} . A_{inc} better reflects the performance of the methods across all the tasks.

4.1. Quantitative Evaluation

We observe that methods proposed for the big-start settings of EFCIL are not effective in small-start settings and perform poorly. A simple baseline trained with LwF and using NCM classifier is performing better than most of the existing approaches - SSRE, PASS, FeTrIL and FeCAM in several settings. While SDC improves over NCM, the proposed method ADC outperforms all existing methods in both last task accuracy and average incremental accuracy across all settings in Tab. 1 and Tab. 2. ADC outperforms the second-best method SDC by 4.2% on 5-task and by 5.12% on 10-task settings of CIFAR-100 on last-task accuracy. For TinyImageNet, ADC improves over the second-best method by 0.95% on 5-task and by 5.17% on 10-task settings. On ImageNet-Subset, ADC is better by 2.58% on 5-task and by 1.72% on 10-task settings after the last task.

We also evaluate the EFCIL methods on the challenging

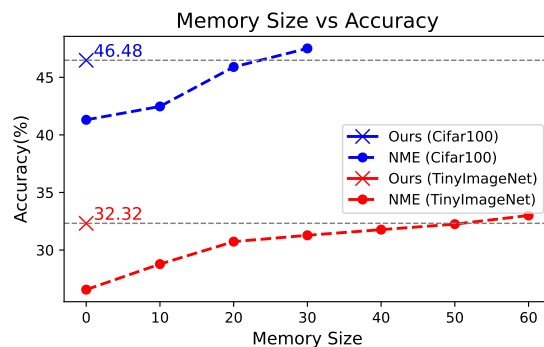


Figure 4. Memory Size vs accuracy comparison of NME and ADC on CIFAR-100 and TinyImageNet (T=10) settings.

fine-grained datasets of CUB-200 and Stanford Cars. We observe in Tab. 2 that LwF is a strong baseline here, particularly in the 5-task and 7-task settings and methods like NCM and SDC are not much better than LwF. While PASS performs poorly on both datasets, FeTrIL and FeCAM performs better with FeCAM outperforming the other methods on the 10-task setting of CUB-200 and 14-task setting of Stanford Cars. ADC outperforms the runner-up methods by 5.78% on 5-task setting and by 6.19% on 10-task settings of CUB-200. On Stanford Cars dataset, ADC is better by 9.68% on 7-task setting and 7.57% on 14-task setting. We analyze how the accuracy after each task varies for all the methods in Fig. 5 and observe that ADC consistently outperforms the other methods across all tasks.

Comparison to NME. We compare the last-task accuracy of ADC with exemplar-based NME where the exemplars are used to estimate the old class prototype positions in the new feature space. We show in Fig. 4 that ADC outperforms NME using 20 exemplars per class for CIFAR-100 (total memory size of 2000 samples) and using 50 exemplars per class (total memory size of 10k samples) for TinyImageNet.

Method	CUB-200				Stanford Cars			
	T=5		T=10		T=7		T=14	
	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}
LwF [24]	<u>58.68</u>	<u>71.31</u>	41.96	60.15	<u>45.18</u>	61.14	30.33	49.93
NCM	52.74	67.13	38.47	57.83	42.22	59.06	31.60	51.34
SDC [52]	55.20	68.64	41.63	60.43	45.03	<u>61.75</u>	32.15	<u>53.18</u>
PASS [55]	34.04	49.00	26.37	41.08	20.71	37.13	12.30	25.46
FeTrIL [35]	54.66	67.45	49.09	62.42	36.92	54.09	34.29	50.41
FeCAM [11]	53.47	66.39	<u>51.78</u>	<u>64.97</u>	40.64	56.24	<u>37.50</u>	52.78
ADC (Ours)	64.46	73.49	57.97	68.91	54.86	67.07	45.07	61.39

Table 2. Evaluation of EFCIL methods on fine-grained datasets. Best results in **bold** and second best results are underlined.

4.2. Computational overhead of ADC

Using ADC requires some additional computation to be made in-between each training session. In this section, we provide an estimation of the additional computation required by our method and compare it to the training time of a single task. At the end of each task, our method requires estimating the drift of each stored prototype (1 per old class) and for each of these, compute several adversarial samples starting from available current task samples. As a consequence, the training time of our method scales linearly with the number of classes. For each class, we compute 100 adversarial samples in a single batch and perform 3 training iterations. In order to perform one iteration, we need to compute the gradient of the adversarial loss with respect to the input image, whose cost is equivalent to the one of a normal training backward pass [41]. So, if we denote the number of classes by N_c , and the number of iterations by N_i , we need to perform $N_c \times N_i$ backward passes. In the case of CIFAR-100 and ImageNet-Subset divided in 10 tasks each containing 10 classes, this means an overhead of, $\sum_{t=1}^9 10 \times t \times 3 = 1350$ backward passes. In contrast, one new task is trained for 100 epochs with a batch size of 128 (39 batches per epochs with 10 tasks on CIFAR-100), which amounts to 3900 backward passes per task, and two times more for the first task (trained for 200 epochs). In total, our method increases the computational cost by 3.1% on this setting. For the 5-task setting of CIFAR-100 and ImageNet-Subset, it increases by 2.5%.

4.3. Ablation Studies

In Sec. 4.3, we conduct an analysis on the impact of various hyperparameters, including the number of iterations, α , and the number of closest samples used for ADC, on CIFAR-100 (T=10) setting. Based on the observations in Tab. 3a, we find that choosing even a very low number of iterations, specifically 3, yields favorable results when generating the perturbed images. Additionally, using $\alpha = 25$ achieves good accuracy for both incremental and final task evalua-

(a) Impact of iterations ($\alpha = 25$)						
Iterations	1	2	3	5	10	
A_{inc}	60.94	61.23	61.35	61.25	60.93	
A_{last}	45.96	46.45	46.48	45.95	45.28	
(b) Impact of α (iterations=3)						
α	1	10	25	50	100	
A_{inc}	60.46	61.14	61.35	60.83	60.93	
A_{last}	44.89	46.43	46.48	45.19	45.28	
(c) Impact of the number of samples						
Samples	25	50	100	300	500	1000
A_{inc}	60.19	60.89	61.35	61.54	61.64	61.47
A_{last}	43.98	45.63	46.48	46.66	46.83	46.32

Table 3. Impact of hyperparameters on CIFAR100 (T=10) setting using the proposed ADC method.

tions in Tab. 3b. Regarding the number of closest samples to the target old prototype, Tab. 3c shows that the accuracy improves marginally on considering more than 100 samples. So, we take the 100 closest samples for our experiments which is computationally cheaper and yet achieves very good accuracy. Interestingly, even taking only the closest 25 samples achieves 2.62% better accuracy than the runner-up method SDC.

Drift estimation quality: We validate through Tab. 1 and Tab. 2 that the designed ADC method is giving better accuracy results than the previous SDC method for all datasets. As an additional verification, we check that this method was indeed better than SDC at estimating the old prototypes drift. To do so, we use both SDC and ADC on the same trained checkpoints on CIFAR-100 5-task settings and compare the estimated drift to the true drift computed using old data. We report the results in Fig. 6, where we show the distribution of the estimated drift qualities. One drift per class is estimated and we compute the cosine similarity of estimated drift to the true drift. We see that for all

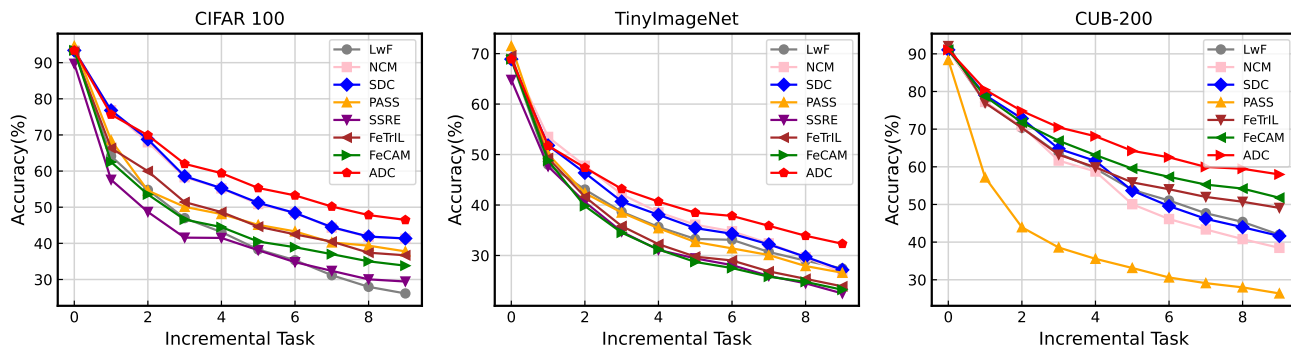


Figure 5. Accuracy after each incremental task for CIFAR-100, TinyImageNet and CUB-200 datasets on 10 task settings. ADC improves over the compared methods starting from the initial to the last task.

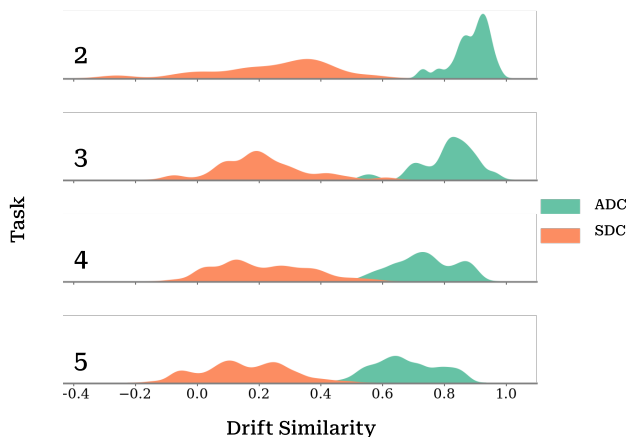


Figure 6. Comparison between the two drift estimation methods SDC [52], and the proposed ADC, on CIFAR-100 (5 tasks). We compute the drift for each class with the two methods and report the distribution of drift estimation quality, measured by computing the cosine similarity between the estimated drift vector and the true drift (obtained using old data), for all previous class prototypes.

training tasks, the drifts estimated with ADC are of better quality than the ones estimated with SDC. We observe that some class drift estimations with SDC have negative cosine similarity with the true drift. However, we also see that the estimation quality decreases slightly for later training tasks. Indeed, as the backbone drifts more and more, it gets harder to estimate the actual drift. The fact that we see this decrease more prominently for ADC might be because the similarities obtained by SDC are already centered around a low-value (0.15) after the second task, whereas the better ADC drift estimation is centered first around 0.9, to then decrease and reach a minimum average of 0.7. This validates that ADC is able to track the movement of the prototypes in the feature space.

5. Conclusions

In this study, we explored a drift compensation method for exemplar-free continual learning. Drawing inspiration

from adversarial attack techniques, we introduced a novel approach called Adversarial Drift Compensation. This method involves generating samples from the new task data in a manner that adversarial images result in embeddings close to the old prototypes. This approach allows us to more accurately estimate the drift of old prototypes in class-incremental learning without the need for any exemplars. Furthermore, we conducted an analysis of continual adversarial transferability, revealing an intriguing observation: generated samples for the old feature space (previous task) continue to behave similarly in the new feature space (current task). This sheds light on why the Adversarial Drift Compensation method performs exceptionally well. Through a series of experiments, we demonstrated that ADC effectively tracks the drift of class distributions in the embedding space, surpassing existing exemplar-free class-incremental learning methods on several standard benchmarks. Importantly, these improvements are achieved without imposing extensive computational overhead or requiring a large memory footprint.

Limitations. The ADC method, as currently designed, requires the access to the task boundaries during training in order to trigger the computation of the old prototypes drift and to access a big enough quantity of current data. The method would for instance be more challenging to use and would require changes in order to be applied in the online continual learning setting, or the continual few-shot learning setting where only a small amount of current data is available. Future work can explore these directions.

Acknowledgement. We acknowledge projects TED2021-132513B-I00 and PID2022-143257NB-I00, financed by MCIN/AEI/10.13039/501100011033 and FSE+ and the Generalitat de Catalunya CERCA Program. This work was partially funded by the European Union under the Horizon Europe Program (HORIZON-CL4-2022-HUMAN-02) under the project “ELIAS: European Lighthouse of AI for Sustainability”, GA no. 101120237. Bartłomiej Twardowski acknowledges the grant RYC2021-032765-I.

References

- [1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 2, 3
- [2] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2
- [3] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, 2018. 1
- [4] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2021. 1, 2
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 5
- [6] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyang Wu, and Rama Chellappa. Learning without memorizing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2
- [7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4
- [8] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2
- [9] Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. Adversarial continual learning. In *European Conference on Computer Vision (ECCV)*, 2020. 3
- [10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICML)*, 2015. 4
- [11] Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 5, 6, 7
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
- [13] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2
- [14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3
- [15] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4
- [16] Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 2
- [17] Xisen Jin, Arka Sadhu, Junyi Du, and Xiang Ren. Gradient-based editing of memory examples for online task-free continual learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3
- [18] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1
- [19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision (ICCV-W) Workshops*, 2013. 2, 5
- [20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5
- [21] Lilly Kumari, Shengjie Wang, Tianyi Zhou, and Jeff A Bilmes. Retrospective adversarial replay for continual learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3
- [22] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2016. 2, 4
- [23] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015. 2, 5
- [24] Zhizhong Li and Derek Hoiem. Learning without forgetting. *Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2017. 2, 5, 6, 7
- [25] Yuyang Liu, Yang Cong, Dipam Goswami, Xialei Liu, and Joost van de Weijer. Augmented box replay: Overcoming foreground shift for incremental object detection. In *International Conference on Computer Vision (ICCV)*, 2023. 1
- [26] Chunwei Ma, Zhanghexuan Ji, Ziyun Huang, Yan Shen, Mingchen Gao, and Jinhui Xu. Progressive voronoi diagram subdivision enables accurate data-free class-incremental learning. In *International Conference on Learning Representations (ICLR)*, 2023. 2
- [27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICML)*, 2018. 2, 4
- [28] Tamasha Malepathirana, Damith Senanayake, and Saman Halgamuge. Napa-vq: Neighborhood-aware prototype augmentation with vector quantization for continual learning. In *International Conference on Computer Vision (ICCV)*, 2023. 2
- [29] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation.

- Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2022. 1, 2, 5
- [30] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Elsevier, 1989. 1, 2
- [31] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4
- [33] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. 3
- [34] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 4
- [35] Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2, 5, 6, 7
- [36] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
- [37] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 1995. 2
- [38] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2
- [39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 2015. 5
- [40] Dawid Rymarczyk, Joost van de Weijer, Bartosz Zieliński, and Bartłomiej Twardowski. Iccle: Interpretable class incremental continual learning. In *International Conference on Computer Vision (ICCV)*, 2023. 5
- [41] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7
- [42] Wuxuan Shi and Mang Ye. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *International Conference on Computer Vision (ICCV)*, 2023. 2
- [43] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *AAAI Conference on Artificial Intelligence*, 2021. 3
- [44] Albin Soutif-Cormerais, Marc Masana, Joost Van de Weijer, and Bartłomiej Twardowski. On the importance of cross-task features for class-incremental learning. *International Conference on Machine Learning (ICML) Workshops*, 2021. 2
- [45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 2, 3
- [46] Marco Toldo and Mete Ozay. Bring evanescent representations to life in lifelong class incremental learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [47] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. 1
- [48] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5
- [49] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European Conference on Computer Vision (ECCV)*, 2022. 1
- [50] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2023. 1
- [51] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deep-inversion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3
- [52] Lu Yu, Bartłomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 5, 6, 7, 8
- [53] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, and De-Chuan Zhan. Pycil: a python toolbox for class-incremental learning. *SCIENCE CHINA Information Sciences*, 2023. 5
- [54] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023. 1, 2
- [55] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 6, 7
- [56] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6
- [57] Kai Zhu, Kecheng Zheng, Ruili Feng, Deli Zhao, Yang Cao, and Zheng-Jun Zha. Self-organizing pathway expansion for non-exemplar class-incremental learning. In *International Conference on Computer Vision (ICCV)*, 2023. 2