

On the Robustness, Generalization, and Forgetting of Shape-Texture Debiased Continual Learning

Zenglin Shi¹, Ying Sun^{1,2}, Joo Hwee Lim^{1,2,3}, Mengmi Zhang^{1,2}

¹I²R and ²CFAR, Agency for Science, Technology and Research, Singapore

³SCSE, Nanyang Technological University, Singapore

Address correspondence to mengmi@i2r.a-star.edu.sg

Abstract

Tremendous progress has been made in continual learning to maintain good performance on old tasks when learning new tasks by tackling the catastrophic forgetting problem of neural networks. This paper advances continual learning by further considering its out-of-distribution robustness, in response to the vulnerability of continually trained models to distribution shifts (e.g., due to data corruptions and domain shifts) in inference. To this end, we propose shape-texture debiased continual learning. The key idea is to learn generalizable and robust representations for each task with shape-texture debiased training. In order to transform standard continual learning to shape-texture debiased continual learning, we propose shape-texture debiased data generation and online shape-texture debiased self-distillation. Experiments on six datasets demonstrate the benefits of our approach in improving generalization and robustness, as well as reducing forgetting. Our analysis on the flatness of the loss landscape explains the advantages. Moreover, our approach can be easily combined with new advanced architectures such as vision transformer, and applied to more challenging scenarios such as exemplar-free continual learning.

1. Introduction

This paper considers the problem of learning a single neural network from a sequence of tasks continually. The continual learning problem is challenging due to *catastrophic forgetting*, which causes an abrupt performance drop on old tasks after learning new tasks [10, 23]. To tackle catastrophic forgetting, regularization-based approaches, e.g., [3, 17, 35], propose to identify and preserve important parameters for old tasks. Knowledge distillation-based approaches, e.g., [4, 7, 21, 26, 33], propose to distill the knowledge of old tasks to the current model by storing the previously trained model. Replay-based approaches,

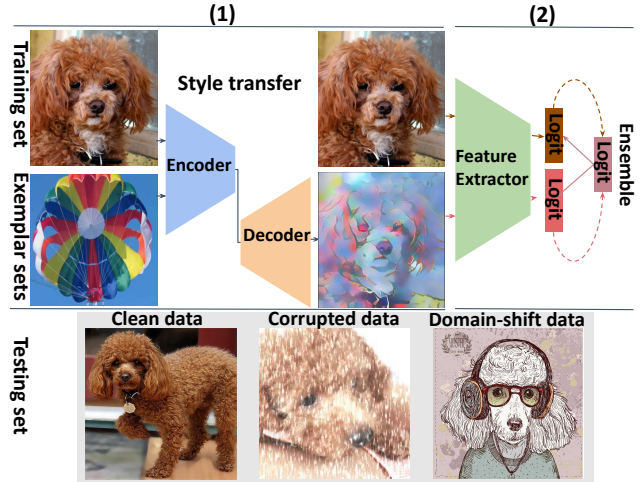


Figure 1. **Shape-texture debiased continual learning.** We introduce two techniques to transform standard continual learning to shape-texture debiased continual learning. (1): *Shape-texture debiased data generation*. Given an image from the training set, we generate its shape-texture conflict image by transferring a new style from an image randomly sampled from the exemplar sets to the image; (2): *Online shape-texture debiased self-distillation*. The original image and shape-texture conflict image are combined in the mini-batch for training. Their output logits are ensemble to get a new soft target for shape-texture debiased distillation. With these two techniques, the continually trained model can generalize to clean data, corrupted data, and domain-shift data well in inference.

e.g., [4, 14, 22, 27, 28, 33], store a few exemplars from old tasks for training alongside new tasks. This paper goes beyond just tackling catastrophic forgetting by learning generalizable and robust representations to further improve generalization and out-of-distribution robustness.

We take inspiration from Li *et al.* [20] and learn generalizable and robust representations by shape-texture debiased training. Shape and texture provide important complementary visual cues for object recognition. This suggests that networks should learn both shape and texture

representations for high-quality recognition. However, Geirhos *et al.* [9] revealed that convolutional networks trained on natural images are biased towards learning texture representations. Texture is sufficient to generalize to unseen in-distribution data that is from the same distribution as the training data. Unfortunately, it often fails to generalize to out-of-distribution data, such as corrupted data and domain-shift data. To improve the model’s out-of-distribution robustness, some works, *e.g.*, [9, 12, 29], propose shape-biased representation learning. These works achieve improved performance on out-of-distribution data, but at the cost of sacrificing the performance on the in-distribution data. To improve the performance on both in-distribution data and out-of-distribution data, Li *et al.* [20] propose shape-texture debiased training to learn both shape and texture representations. However, their method is not specifically catered for continual learning, and underperforms as shown by our experiments. Instead, this paper introduces the idea of shape-texture debiased training to continual learning scenarios by specifically proposing a shape-texture debiased continual learning algorithm.

We make the following contributions. (I) we investigate and improve the out-of-distribution robustness for the first time in continual learning, in addition to considering generalization and forgetting; (II) we propose shape-texture debiased continual learning to learn generalizable and robust representations continually with shape-texture debiased data generation and online shape-texture debiased self-distillation, as illustrated in **Figure 1**; (III) we provide extensive experiments to demonstrate the benefits of our proposed approach in improving generalization and robustness, as well as reducing forgetting. Moreover, we show our approach has flatter local minima than standard continual learning, which explains its advantages. We also connect our approach with vision transformer and generalize our approach to exemplar-free continual learning.

2. Related work

2.1. Continual learning

Continual learning is to learn a single neural network from a sequence of tasks incrementally, in which only data from the current task is available for training. A significant challenge in continual learning is that neural networks tend to exhibit high plasticity on new tasks, but suffer from low stability on previously learned tasks. High plasticity allows neural networks to learn new knowledge and generalize to new tasks quickly, while low stability causes catastrophic forgetting of the knowledge acquired from old tasks [10, 23]. Due to catastrophic forgetting, the performance on old tasks drops drastically after learning new tasks.

In order to tackle the catastrophic forgetting problem, regularization-based approaches, *e.g.*, [3, 17, 35], try to

impose strong constraints on model parameters by penalizing the changes of parameters that are important for old tasks. The difference between these approaches lies in the way to measure the importance. While these approaches have shown to mitigate forgetting to some extent, they fail to achieve satisfactory performance in challenging continual learning scenarios [31, 33]. Alternatively, knowledge distillation [13] has been recently used as a regularizer to minimize the changes to the decision boundaries of old tasks while learning new tasks. Knowledge distillation-based approaches, *e.g.*, [4, 7, 21, 26, 33], usually store a snapshot of the model learned from old tasks in a memory buffer, and then distill the knowledge encoded in the stored old model to the current model. Another line of approaches mitigating catastrophic forgetting considers exemplars replay, *e.g.*, [4, 14, 22, 27, 28, 33]. Replay-based approaches store a few exemplars from old tasks in a memory buffer, and many of them, *e.g.*, [4, 14, 27, 33] use herding heuristics [32] for exemplar selection. The stored exemplars combined with the new data are used to optimize the network parameters when learning a new task.

In this work, we go beyond just solving the catastrophic forgetting by learning robust and generalized representations for each task with shape-texture debiased training. Moreover, we show that shape-texture debiased training acts as an implicit regularization to mitigate forgetting.

2.2. Shape- & texture-biased learning

Geirhos *et al.* [9] show that, unlike humans, convolutional networks tend to classify images by texture rather than shape. Although texture alone is sufficient to generalize to unseen test data from the same distribution as the training data, it often fails to generalize to out-of-distribution data. To improve the model’s robustness against out-of-distribution data, some methods force the model to be less reliant on texture information. For example, Geirhos *et al.* [9] proposed to change the textures of training data with new textures from artistic paintings to decrease the texture bias in favor of increased shape bias. Hermann *et al.* [12] showed that data augmentation is another effective way to reduce the model’s texture bias. Instead of augmentation, Shi *et al.* [29] alternatively propose an informative dropout filter to explicitly capture shape information.

However, these shape-biased representation learning methods sacrifice the accuracy on the in-distribution dataset while achieving improved performance on the out-of-distribution dataset. In order to address this issue, Li *et al.* [20] propose shape-texture debiased training to avoid being biased towards learning either shape or texture representations. They augment the original training data using style transfer to break the correlation between label and texture as well. The critical thing is that they provide the corresponding supervisions from both shape

and texture to learn shape-texture representations at the same time. The shape-texture debiased learning improves the out-of-distribution robustness without losing accuracy on in-distribution data. However, their approach is not specifically developed for continual learning. Simply applying their approach to the continual learning scenarios, therefore, is inferior, as demonstrated by our experiments.

In this work, we propose shape-texture debiased data generation and online shape-texture debiased self-distillation specifically for developing shape-texture debiased continual learning, leading to superior results.

3. Method

3.1. Standard continual learning setup

In this work, we focus on the class-incremental scenario in continual learning. The goal is to learn a unified classifier over incrementally occurring sets of classes [27]. Formally, let $\mathcal{T} = \{\mathcal{T}_0, \dots, \mathcal{T}_T\}$ be a sequence of classification tasks, $\mathcal{D}_t = \{(x_{i,t}, y_{i,t})\}_{i=0}^{n_t}$ be the labeled training set of the task \mathcal{T}_t with n_t samples from m_t distinct classes, and $(x_{i,t}, y_{i,t})$ be the i -th (image, label) pair in the training set of the task \mathcal{T}_t . A single model Θ is trained to solve the tasks \mathcal{T} incrementally. The model Θ consists of a feature extractor and a unified classifier. For the initial task \mathcal{T}_0 , the model Θ_0 learns a standard classifier for the first m_0 classes. In the incremental step t , only the classifier is extended by adding m_t new output nodes to learn m_t new classes, leading to a new model Θ_t .

It's known that naively training the model Θ over the sequence of data $\{\mathcal{D}_0, \dots, \mathcal{D}_T\}$ causes catastrophic forgetting. Following recent works, e.g., [2, 7, 8, 14, 22, 26], exemplars replay and knowledge distillation are used to handle catastrophic forgetting. Therefore, at each incremental step, a few exemplars \mathcal{P}_t are retained in a memory buffer for replay. Meanwhile, the snapshot of the parameters of the model Θ_t is retained in the memory buffer for knowledge distillation. The overall loss function at each incremental step t is defined as:

$$\mathcal{L}_t = \mathcal{L}_{\mathcal{D}_t}^{CE} + \mathcal{L}_{\mathcal{P}}^{CE} + \lambda * (\mathcal{L}_{\mathcal{D}_t}^{KD} + \mathcal{L}_{\mathcal{P}}^{KD}), \quad (1)$$

where cross-entropy loss $\mathcal{L}_{\mathcal{D}_t}^{CE}$ and knowledge distillation loss $\mathcal{L}_{\mathcal{D}_t}^{KD}$ are computed on the current training set \mathcal{D}_t for learning the new classes. For replay, $\mathcal{L}_{\mathcal{P}}^{CE}$ and $\mathcal{L}_{\mathcal{P}}^{KD}$ are computed on the exemplar sets $\mathcal{P} = \{\mathcal{P}_0, \dots, \mathcal{P}_t\}$. λ is a hyper-parameter to control the strength of knowledge distillation loss. So far we have introduced the general setup for standard continual learning. Next, we elaborate on how to transform the standard continual learning to shape-texture debiased continual learning with two main components, i.e., shape-texture debiased data generation and online shape-texture debiased self-distillation.

3.2. Shape-texture debiased data generation

Our starting point is to improve out-of-distribution robustness in continual learning, in response to the vulnerability of continually trained models to the distribution shifts (e.g., data corruptions and domain shifts). The standard training on the natural images tends to learn texture-biased representations. However, shape-biased representations are more robust against out-of-distribution than texture-biased representations [9]. To reduce texture-bias and increase shape-bias, we propose to generate shape-texture conflict images for shape-texture debiased training.

Inspired by Geirhos *et al.* [9], we generate a shape-texture conflict image by transferring a new style to the original natural image. *The biggest challenge faced in the generation process is how to obtain diverse new styles.* Geirhos *et al.* [9] obtain new styles from artistic paintings. This is not suitable for memory-constrained continual learning because of the extra memory required to store the paintings. Moreover, the significant domain gap between natural images and artistic paintings results in generated images with low quality. An alternative solution is to use the styles from different images in the current training set. However, the images in the current training set are usually from a few classes in each incremental step. Thus, the styles from the images in the current training set are not diverse. To alleviate this issue, we instead propose to use the styles from the images in exemplar sets. The images in exemplar sets are already stored for replay and are natural without domain conflict. More importantly, the images in exemplar sets are from much more classes, introducing more diverse styles. We also propose to augment styles by style distortions to increase the diversity of styles further.

Let $\mathcal{X} = \{x_{1,t}, \dots, x_{k,t}\}$ be a mini-batch of images with a size of k randomly sampled from the current training set \mathcal{D}_t at the incremental step t , and $\hat{\mathcal{X}} = \{\hat{x}_1, \dots, \hat{x}_k\}$ be a mini-batch of images with a size of k randomly sampled from the exemplar sets $\mathcal{P} = \{\mathcal{P}_0, \dots, \mathcal{P}_t\}$. Then the mini-batch of shape-texture conflict images $\tilde{\mathcal{X}} = \{\tilde{x}_1^t, \dots, \tilde{x}_k^t\}$ can be generated based on \mathcal{X} and $\hat{\mathcal{X}}$, in which

$$\tilde{x}_{i,t} = f(x_{i,t}, \phi(\hat{x}_i)). \quad (2)$$

Here $f(\cdot)$ is the style transfer operation and $\phi(\cdot)$ is the style distortion operation. In this paper, we use the real-time style transfer approach, AdaIN [15]. For style distortion, we use *ColorJitter*, *RandomGrayscale*, *RandomGaussianBlur*, and *RandomGaussianNoise*. We avoid using style distortions directly on the image $x_{i,t}$ because distortions make it difficult for f to extract the shape information, leading to shape-texture conflict images with low quality.

3.3. Online shape-texture debiased self-distillation

Training only on the original natural images results in learning texture-biased representations, while training only

on shape-texture conflict images tends to learn shape-biased representations. To avoid representation learning being overly biased towards shape or texture, we propose to train the models jointly on both shape-texture conflict images and original natural images. To further enhance the debiased training, we introduce an online shape-texture debiased self-distillation.

Let $\mathcal{X} = \{x_{1,t}, \dots, x_{k,t}\}$ be a mini-batch of images with a size of k randomly sampled from the current training set \mathcal{D}_t in the incremental step t , and $\tilde{\mathcal{X}} = \{\tilde{x}_{1,t}, \dots, \tilde{x}_{k,t}\}$ be the corresponding mini-batch of shape-texture conflict images generated following Eq. (2). Then we obtain a new mini-batch $\{\mathcal{X}, \tilde{\mathcal{X}}\}$ by combining \mathcal{X} and $\tilde{\mathcal{X}}$ for model training. Let $\mathcal{Z} = \{z_{1,t}, \dots, z_{k,t}\}$ and $\tilde{\mathcal{Z}} = \{\tilde{z}_{1,t}, \dots, \tilde{z}_{k,t}\}$ be the corresponding logit of \mathcal{X} and $\tilde{\mathcal{X}}$, produced by the model Θ_t , respectively. Intuitively, \mathcal{Z} encodes texture information while $\tilde{\mathcal{Z}}$ encodes shape information. The shape-texture information is then ensembled to form a better soft target for distillation. The distillation loss is defined by

$$\mathcal{L}^{STD} = D_{KL}(p||q) + D_{KL}(\tilde{p}||q), \quad (3)$$

where $p = \sigma(\mathcal{Z}/\tau)$, $\tilde{p} = \sigma(\tilde{\mathcal{Z}}/\tau)$, $q = \sigma((\mathcal{Z} + \tilde{\mathcal{Z}})/2\tau)$. τ is a temperature hyper-parameter, $\sigma(\cdot)$ is a softmax function, and $(\mathcal{Z} + \tilde{\mathcal{Z}})/2$ is the ensembled logit. Our distillation generates a soft target that integrates shape and texture knowledge from a mini-batch in an online manner, which only requires minimal computations without the need of introducing extra network parameters.

3.4. Shape-texture debiased continual learning

Algorithm 1 describes our shape-texture debiased continual learning. We train the models for each task incrementally on both shape-texture conflict images and original natural images. In addition to using classification loss and knowledge distillation loss as defined in Eq. 1, we also employ our online shape-texture debiased self-distillation loss. We don't perform shape-texture debiased training in replay because some shape-texture conflict images may be synthesized with low quality, introducing noises. Replay only with the original natural images from exemplar sets is in favor of reducing the harmful effects of noises and improving performance. The overall loss function at each incremental step t is defined as:

$$\mathcal{L}_t = \mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{CE} + \mathcal{L}_{\mathcal{P}}^{CE} + \lambda * (\mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{KD} + \mathcal{L}_{\mathcal{P}}^{KD}) + \gamma \mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{STD}, \quad (4)$$

where cross-entropy loss $\mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{CE}$ and knowledge distillation loss $\mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{KD}$ are computed on the combination of shape-texture conflict images \mathcal{D}_t and the original natural images $\tilde{\mathcal{D}}_t$ from the current task t . For replay, $\mathcal{L}_{\mathcal{P}}^{CE}$ and $\mathcal{L}_{\mathcal{P}}^{KD}$ are computed only on the original natural images from exemplar sets \mathcal{P} . The last item is the proposed online shape-texture debiased self-distillation loss. λ and

Algorithm 1: Shape-texture debiased continual learning

input: \mathcal{D}_t ; /*the training set for the new task */
require: $\mathcal{P} = \{\mathcal{P}_i\}_{i=0}^{t-1}$; /*the current exemplar sets for replay*/
require: Θ_{t-1} ; /*the old model learned from the old task*/
require: f ; /*the style transfer model*/
output: Θ_t ; /*the new model to be learned from the new task*/

- 1 $\Theta_t \leftarrow \Theta_{t-1}$; /*add m_t new output nodes*/
- 2 construct new exemplar set \mathcal{P}_t ;
- 3 $\mathcal{P} \leftarrow \{\mathcal{P}_i\}_{i=0}^{t-1} \cup \mathcal{P}_t$; /*add \mathcal{P}_t to replay buffer*/
- 4 **for** iteration $j \leftarrow 1$ **to** l **do**
- 5 load a mini-batch $(\mathcal{X}_t, \mathcal{Y}_t)$ from \mathcal{D}_t ;
- 6 $\mathcal{X}_t \leftarrow \{(x_{i,t})\}_{i=1}^k$; $\mathcal{Y}_t \leftarrow \{(y_{i,t})\}_{i=1}^k$;
- 7 load a mini-batch $(\hat{\mathcal{X}}_t, \hat{\mathcal{Y}}_t)$ from \mathcal{P} ;
- 8 $\hat{\mathcal{X}}_t \leftarrow \{(\hat{x}_{i,t})\}_{i=1}^k$; $\hat{\mathcal{Y}}_t \leftarrow \{(\hat{y}_{i,t})\}_{i=1}^k$;
- 9 $\tilde{\mathcal{X}}_t \leftarrow \{f(x_{i,t}, \phi(\hat{x}_i))\}_{i=1}^k$; /*using Eq.(2)*/
- 10 $\mathcal{Z}_t, \tilde{\mathcal{Z}}_t \leftarrow \Theta_t(\mathcal{X}_t, \tilde{\mathcal{X}}_t)$; /*forward pass for new task*/
- 11 Compute $\mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{STD}$ with $(\mathcal{Z}_t, \tilde{\mathcal{Z}}_t)$ using Eq.(3);
- 12 Compute $\mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{CE}$ with $\{(\mathcal{Z}_t, \mathcal{Y}_t), (\tilde{\mathcal{Z}}_t, \mathcal{Y}_t)\}$;
- 13 $\mathcal{Z}_{t-1}, \tilde{\mathcal{Z}}_{t-1} \leftarrow \Theta_{t-1}(\mathcal{X}_t, \tilde{\mathcal{X}}_t)$; /*distillation*/
- 14 Compute $\mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{KD}$ with $\{(\mathcal{Z}_t, \mathcal{Z}_{t-1}), (\tilde{\mathcal{Z}}_t, \tilde{\mathcal{Z}}_{t-1})\}$
- 15 $\hat{\mathcal{Z}}_t \leftarrow \Theta_t(\hat{\mathcal{X}}_t)$; /*forward pass for replay*/
- 16 Compute $\mathcal{L}_{\mathcal{P}}^{CE}$ with $(\hat{\mathcal{Z}}_t, \hat{\mathcal{Y}}_t)$;
- 17 $\hat{\mathcal{Z}}_{t-1} \leftarrow \Theta_{t-1}(\hat{\mathcal{X}}_t)$; /*distillation for replay*/
- 18 Compute $\mathcal{L}_{\mathcal{P}}^{KD}$ with $(\hat{\mathcal{Z}}_t, \hat{\mathcal{Z}}_{t-1})$;
- 19 $\mathcal{L}_t = \mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{CE} + \mathcal{L}_{\mathcal{P}}^{CE} + \lambda * (\mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{KD} + \mathcal{L}_{\mathcal{P}}^{KD}) + \gamma \mathcal{L}_{\mathcal{D}_t, \tilde{\mathcal{D}}_t}^{STD}$
 optimize Θ_t with loss \mathcal{L}_t ; /*backward pass*/

γ are two hyper-parameters to control the strength of knowledge distillation loss and online shape-texture debiased self-distillation loss.

4. Experiments and results

4.1. Experimental setup

Datasets. We perform class-incremental experiments on three standard datasets including CIFAR-100 [18], ImageNet-100, and ImageNet-1000 [5], following, e.g., [8, 14, 22, 30]. CIFAR-100 contains 60,000 color images with the same size of 32×32 from 100 classes, in which 50,000 images for training and the remaining 10,000 for testing. ImageNet-1000 contains around 1.3 million color images with different sizes from 1,000 classes. For each class, there are 1,300 images for training and 50 images for validation. ImageNet-100 is a subset of Imagenet-1000 with 100 classes that are randomly sampled using the NumPy seed of 1993. We use the dataset ImageNet-1000-C [1], CIFAR-100-C [1], and ImageNet-1000-R [11] for evaluating out-of-distribution robustness. ImageNet-1000-C and CIFAR-100-C consists of 15 diverse corruption types applied to validation images

	ImageNet-100				ImageNet-1000				CIFAR-100		
	$R-C\downarrow$	$R-R\uparrow$	Avg. Acc. \uparrow	$\mathcal{F}\downarrow$	$R-C\downarrow$	$R-R\uparrow$	Avg. Acc. \uparrow	$\mathcal{F}\downarrow$	$R-C\downarrow$	Avg. Acc. \uparrow	$\mathcal{F}\downarrow$
$T=5$											
Standard	86.02	25.8	77.03	12.88	74.19	15.64	67.46	12.94	67.38	64.33	18.36
Ours	69.74	40.12	79.16	7.80	69.12	19.66	67.84	10.92	62.81	65.13	13.64
$T=10$											
Standard	92.96	23.71	74.25	16.12	76.31	14.31	65.20	17.46	70.07	63.01	21.06
Ours	75.45	34.58	76.72	12.60	71.92	17.84	66.47	14.31	63.83	63.69	15.94

Table 1. **Overall performance analysis of shape-texture debiased continual learning.** Compared to standard continual learning (Standard), shape-texture debiased continual learning (Ours) improves generalization (Avg. Acc. \uparrow), reduces forgetting (\mathcal{F}), and boosts robustness against out-of-distribution ($R-C$ and $R-R$) considerably across protocols and datasets.

of ImageNet and CIFAR-100, respectively. There are five different severity levels for each corruption type. ImageNet-1000-R is a real-world distribution shift dataset consisting of changes in image style, image blurriness, geographic location, camera operation, and more, with a total of 30,000 images from 200 ImageNet classes. Following the same way to get ImageNet-100 from ImageNet-1000, we obtain ImageNet-100-C from ImageNet-1000-C and ImageNet-100-R from ImageNet-1000-R.

Protocols. To mimic the practical class-incremental learning, we adopt the protocol proposed in [14] originally, and followed by many works, *e.g.*, [7, 8, 14, 22, 26, 30]. The protocol starts from an initial base task \mathcal{T}_0 followed by T incremental tasks. T is set to 5 and 10. For all datasets, half of the classes are used for the initial base task, and the remaining classes are equally divided over the incremental steps. A strict memory budget is also considered, where only 20 exemplars for each class are saved for replay.

Metrics. We report the experimental results with five metrics. Specifically, *average incremental accuracy* [27], denoted by **Avg. Acc.**, is computed by averaging the accuracies of the models $\{\Theta_0, \dots, \Theta_T\}$ obtained in incremental steps. The accuracy of each model Θ_i obtained in each incremental step is evaluated on all classes seen so far. *Forgetting rate* [22], denoted by \mathcal{F} , measures the performance drop on the initial base task \mathcal{T}_0 . It is the difference between the accuracy of Θ_0 and the accuracy of Θ_T on the same testing data of \mathcal{T}_0 . *The accuracy of the final model* Θ_T is denoted by **Acc.** The robustness of the final model Θ_T against data corruptions is measured by computing mean corruption error on ImageNet-100-C, ImageNet-1000-C and CIFAR-100-C [1], denoted by **$R-C$** . The robustness of the final model Θ_T against domain shifts is measured by computing accuracy on ImageNet-100-R and ImageNet-1000-R, denoted by **$R-R$** .

Implementation details. The proposed approach is implemented based on the publicly available official code provided by Mittal *et al.* [26] using NVIDIA RTX A5000 GPUs. For ImageNet-100 and ImageNet-1000, we use an 18-layer ResNet. The network is trained for 70 epochs in

the initial base task with a base learning rate of $1e-1$, and is trained for 40 epochs in the incremental task with a base learning rate of $1e-2$. The base learning rate is divided by 10 at epochs $\{30, 60\}$ in the initial base task, and is divided by 10 at epochs $\{25, 35\}$ in the incremental steps. For CIFAR-100, we use a 32-layer ResNet. The network is trained for 120 epochs in the initial base task with a base learning rate of $1e-1$, and is trained for 60 epochs in the subsequent incremental tasks with a base learning rate of $1e-2$. We adopt a cosine learning rate schedule, in which the learning rate is decayed until $1e-4$. For all networks, the last layer is cosine normalized, as suggested in [14]. All networks are optimized using the SGD optimizer with a mini-batch of 128, a momentum of 0.9, and a weight decay of $1e-4$. Following [14, 26], we use an adaptive weighting function for λ in **Eq. (4)**. For each incremental step t , $\lambda_t = \lambda_{base} \left(\sum_{i=0}^t m_i / m_t \right)^{2/3}$ where $\sum_{i=0}^t m_i$ denotes the number of all seen classes from step 0 to step t , and m_t denotes the number of classes in step t . λ_{base} is set to 20 for CIFAR-100, 100 for ImageNet-100, and 600 for ImageNet-1000, as suggested by [26]. γ is set to 0.01 by default. All codes will be made publicly available.

4.2. Overall performance analysis

We first analyze the benefits of our shape-texture debiased continual learning over the standard continual learning, with regard to robustness, generalization, and forgetting. For fair comparison, both our shape-texture debiased continual learning and the standard continual learning are implemented based on the publicly available official code provided by Mittal *et al.* [26]. Our approach is simply implemented by adding the proposed shape-texture debiased data generation and online shape-texture debiased self-distillation to the standard one. In **Table 1**, we provide results based on two protocols and three datasets. The results across all experiments are consistent: our shape-texture debiased continual learning has clear advantages over standard continual learning in terms of generalization, forgetting, and robustness. For example, incrementally learning 5 classification tasks on ImageNet-100, our approach improves

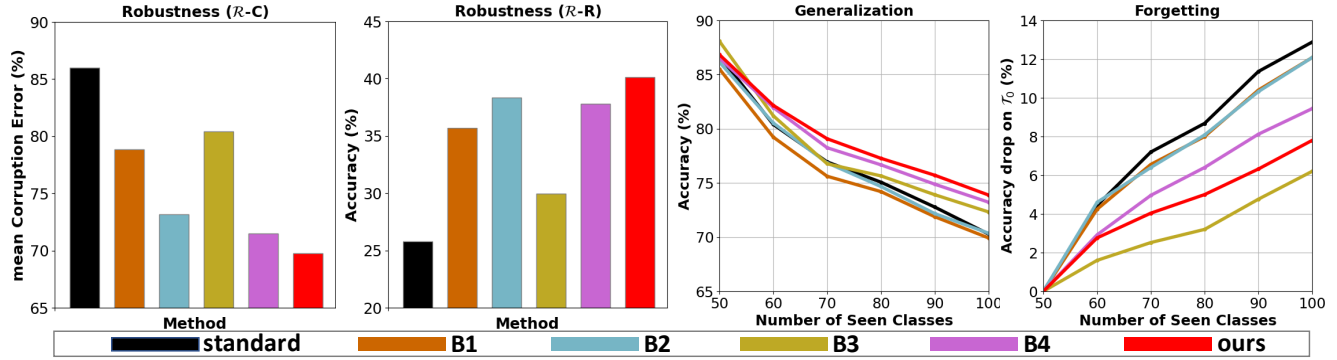


Figure 2. **Ablation study and comparison.** on ImageNet-100 with two protocols. The effectiveness of shape-texture debiased data generation (Section 3.2) and online shape-texture debiased self-distillation (Section 3.3) is demonstrated by the improvements made by **B4** and **ours** over standard continual learning (*standard*), respectively. Also, our approaches achieve the best performance in improving robustness and generalization when compared to the baselines.

the Avg. Acc. from 77.03 to 79.16, and reduces the \mathcal{F} from 12.88 to 7.8. Especially, our approach reduces R -C by 16.29 and improves R -R by 14.32. From this experiment, we come to the conclusion that the generalization, robustness, and forgetting of continual learning can be jointly improved by shape-texture debiased training.

4.3. Comparison to baselines

Considering there exists no prior work dealing with out-of-distribution robustness in continual learning, we implement the following set of baselines to further evaluate the benefits of the proposed shape-texture debiased continual learning. **B1:** We follow Hermann *et al.* [12] to increase shape-bias and reduce texture-bias by applying simple, naturalistic augmentation (color distortion, noise, and blur). Specifically, we use the color jitter with probability of 80%, the color drop with probability of 20%, the Gaussian noise (mean=0 and std=0.025) with probability of 50%, and the Gaussian blur (kernel size was 10% of the image width/height) with probability of 50%, as specified by [12]. **B2:** We follow Geirhos *et al.* [9] to generate shape-texture conflict images using the styles from artistic paintings, which requires additional memory to store the paintings. This baseline trains the models jointly on both shape-texture conflict images and original natural images. **B3:** We follow Li *et al.* [20] to generate shape-texture conflict images by style transfer using AdaIN, where the used styles are from the current training set. This baseline also trains the models jointly on both shape-texture conflict images and original natural images, but additionally provides a mixup loss (refer to Eq. (1) in [20] and λ is set to 0.8 as suggested) for shape-texture conflict images. **B4:** We generate shape-texture conflict images using the proposed approach in Section 3.2 and train the models jointly on both shape-texture conflict images and original natural images. **Standard:** Here we also provide the results of standard

continual learning for analysis.

We provide results on ImageNet-100 with the protocol of $T=5$ in Figure 2. We make the following main observations: 1) Compared to the **standard**, **B1**, and **B2**, **B4** makes improvements across metrics. This demonstrates the effectiveness of our shape-texture debiased data generalization. **ours** further performs better than **B4**, which demonstrates the effectiveness of our online shape-texture debiased self-distillation; 2) In terms of robustness, all methods outperform the **standard**, among that our approaches perform best; 3) In terms of generalization, our approaches make the most improvements, while **B1** and **B2** degrade the performance when compared to the **standard**; 4) In terms of forgetting, all methods outperform the *standard*. This suggests that training with shape-texture conflict images acts as an implicit regularization to mitigate forgetting. **B3** performs best to reduce forgetting, while our approaches outperform **B3** with regard to robustness and generalization improvements.

4.4. Discussions and further analysis

Analysis on the flatness of local minima. Mirzadeh *et al.* [25] revealed that each task in continual learning should push its learning towards flatter local minima to improve generalization and reduce forgetting. Here, we visualize the weight loss landscape of our shape-texture debiased continual learning. We use the visualization approach proposed by Li *et al.* [19]. We plot the changes in its training loss when moving the weights of the network in a random direction with a magnitude. Considering the direction is randomly selected, the visualization is performed using the average result of five different directions.

We first study the forgetting of the networks by plotting changes in the weight loss landscape for \mathcal{T}_0 after \mathcal{T}_5 . As shown in Figure 3a, our approach has less forgetting than the standard one, corresponding to a flatter local minimum.

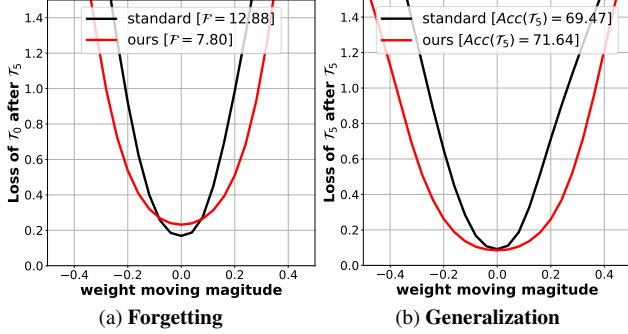


Figure 3. **Analysis on the flatness of local minima.** (a) is the weight loss landscape of the initial task T_0 after learning the fifth task T_5 . Our approach has flatter local minima than standard continual learning, which explains why our approach is able to reduce forgetting from $\mathcal{F} = 12.88$ to $\mathcal{F} = 7.80$; (b) is weight loss landscape of the fifth task T_5 after just learning the fifth task T_5 . Our approach improves the accuracy of T_5 from 69.47 to 71.64 by widening the local minima.

We further evaluate the generalization of the networks by observing the performance of T_5 just after it has been learned. As shown in Figure 3b, our approach shows higher accuracy with a flatter loss minimum. Thus, these results explain why our shape-texture biased continual learning can improve generalization and reduce forgetting.

Styles from exemplar sets vs. current training set. We introduced the style transfer approach to generate shape-texture conflict images in Section 3.2. A key problem we have solved is how to obtain diverse new styles. We proposed to use the styles from the exemplar sets instead of the current training set. This is because exemplar sets provide more diverse styles than the current training set. Here, we perform an empirical comparison. According to the results in Table 2, using the styles from exemplar sets indeed achieves better performance than using the styles from current data across metrics.

	$R-C \downarrow$	$R-R \uparrow$	Avg. Acc. \uparrow	$\mathcal{F} \downarrow$
Current training set	71.70	38.83	78.86	9.44
Exemplar sets	69.74	40.12	79.16	7.80

Table 2. **Styles from exemplar sets vs. current training set.** on ImageNet-100 with the protocol of $T=5$. For shape-texture conflict image generation, using the styles from exemplar sets performs better than using the styles from current data across metrics.

Shape-texture debiased learning in replay? We analyze the effect of shape-texture debiased learning in replay. As explained in Section 3.4, fine-tuning only with the original natural images in exemplar sets may reduce the harmful effect of the noises from shape-texture conflict images. The results in Table 3 indeed show performing shape-texture debiased learning in replay is unfavorable.

Sensitivity analysis on coefficient γ . There is a coefficient

	$R-C \downarrow$	$R-R \uparrow$	Avg. Acc. \uparrow	$\mathcal{F} \downarrow$
Yes	71.68	38.45	78.58	8.91
No	69.74	40.12	79.16	7.80

Table 3. **Effect of shape-texture debiased learning in replay** on ImageNet-100 with the protocol of $T=5$. Performing shape-texture debiased learning in replay is unfavorable.

γ in the overall loss function (see Eq. 4) to control the strength of the proposed online shape-texture debiased self-distillation loss. Here we empirically study five settings, *i.e.*, $\gamma \in \{0, 0.001, 0.01, 0.1, 1\}$. $\gamma=0$ means that we don’t use the loss. The results are shown in Table 4. We observe that using the self-distillation loss with $\gamma \in \{0.001, 0.01, 0.1, 1\}$ performs better than $\gamma=0$ across metrics. This indicates the importance of the self-distillation loss for improving performance. We also observe that higher values lead to less forgetting, but may deteriorate generalization and robustness. Setting a suitable value (*e.g.*, $\gamma=0.01$) results in a good overall performance with regard to all metrics.

	$R-C \downarrow$	$R-R \uparrow$	Avg. Acc. \uparrow	$\mathcal{F} \downarrow$
$\gamma = 0$	71.5	37.8	78.58	9.44
$\gamma = 0.001$	71.36	39.12	78.97	8.32
$\gamma = 0.01$	69.74	40.12	79.16	7.80
$\gamma = 0.1$	70.98	39.19	78.89	7.12
$\gamma = 1$	69.79	39.12	78.73	6.48

Table 4. **Sensitivity analysis on coefficient γ** on ImageNet-100 with the protocol of $T=5$. Setting a suitable value (*e.g.*, $\gamma=0.01$) results in a good overall performance.

Generalize to exemplar-free scenario. We analyze the effect of our shape-texture debiased continual learning for the challenging exemplar-free scenario, in which exemplars are not stored anymore for replay. Note that we still keep knowledge distillation to tackle catastrophic forgetting. We generate shape-texture conflict images using the styles from the current training set instead of exemplar sets. The results are shown in Table 6. Our approach is still superior to the standard continual learning in the exemplar-free scenario.

Connect to vision transformer. We analyze the effect of our shape-texture debiased continual learning with newly emerged powerful vision transformer [6]. Yu *et al.* [34] propose a working recipe of using vision transformer in continual learning. We follow their recipe, but replace their standard training with our shape-texture debiased training. We implement the recipe of Yu *et al.* [34] following their paper. The results are shown in Table 7. Our shape-texture debiased continual learning still has superior performance over standard continual learning with vision transformer across metrics. This also suggests that our approach is architecture-agnostic.

	ImageNet-100			ImageNet-1000			CIFAR-100		
	Avg. Acc. \uparrow	Acc. \uparrow	$\mathcal{F} \downarrow$	Avg. Acc. \uparrow	Acc. \uparrow	$\mathcal{F} \downarrow$	Avg. Acc. \uparrow	Acc. \uparrow	$\mathcal{F} \downarrow$
$T=5$									
LUCIR [14]	70.84 \pm 0.69	60.00	31.88	64.45	56.60	24.08	63.17 \pm 0.87	54.30	18.70
CLFAS [24]	75.85	-	-	-	-	-	65.44	-	-
Mnemonics [22]	75.54 \pm 0.85	61.36	17.40	64.54	56.85	13.85	63.34 \pm 0.34	54.32	10.91
Podnet [7]	76.96 \pm 0.29	67.60	-	66.43	58.90	-	64.83 \pm 1.11	54.60	-
CCIL* [26]	77.03 \pm 0.51	70.28	12.88	67.46	60.83	12.94	64.33 \pm 0.77	56.82	18.36
AFC [16]	76.87	-	-	68.90	-	-	66.49 \pm 0.81	-	-
Ours	79.16 \pm 0.43	73.88	7.80	67.84	61.62	10.92	65.13 \pm 0.69	57.35	13.64
$T=10$									
LUCIR [14]	68.32 \pm 0.81	57.10	33.48	61.57	51.70	27.29	60.14 \pm 0.73	50.30	21.34
CLFAS [24]	72.11	-	-	-	-	-	62.48	-	-
Mnemonics [22]	74.33 \pm 0.56	59.56	17.08	63.01	54.99	15.82	62.28 \pm 0.61	51.53	13.38
Podnet [7]	73.70 \pm 1.05	65.00	-	63.21	55.70	-	63.19 \pm 1.31	53.00	-
CCIL* [26]	74.25 \pm 0.67	64.80	16.12	65.20	57.13	17.46	63.01 \pm 0.81	53.02	21.06
AFC [16]	75.75	-	-	67.02	-	-	64.98 \pm 0.87	-	-
Ours	76.72 \pm 0.58	69.16	12.60	66.47	58.49	14.31	63.69 \pm 0.77	53.98	15.94

Table 5. **Comparison to the state-of-the-art** on ImageNet-100, ImageNet-1000, and CIFAR-100. The results obtained by our approach are either competitive or better than the current state-of-the-art across datasets, protocols, and metrics. * based on our re-implementation.

	$R-C \downarrow$	$R-R \uparrow$	Avg. Acc. \uparrow	$\mathcal{F} \downarrow$
Standard	92.01	22.81	72.33	4.08
Ours	79.44	33.52	75.61	3.44

Table 6. **Generalize to exemplar-free scenario** on ImageNet-100 with the protocol of $T=5$. In the challenging exemplar-free scenario, Our approach is still superior to the standard ones.

	$R-C \downarrow$	$R-R \uparrow$	Avg. Acc. \uparrow	$\mathcal{F} \downarrow$
Yu <i>et al.</i> [34]	82.36	24.54	78.17	6.56
Ours	76.75	28.42	78.71	5.12

Table 7. **Connecting to vision transformer** on ImageNet-100 with the protocol of $T=5$. Replacing convolutional networks with the advanced vision transformer, our shape-texture debiased continual learning still has superior performance over standard continual learning (Yu *et al.* [34]) across all metrics.

4.5. Comparison to the state-of-the-art.

Finally, we draw a comparison to the state-of-the-art in class-incremental learning based on three datasets and two protocols. We compare with six recent approaches, including LUCIR [14], CLFAS [24], Mnemonics [22], Podnet [7], AFC [16], and CCIL [26]. Since the first five approaches report their results using the same protocols to conduct class-incremental learning as ours, we directly use their reported results for comparison. Considering CCIL does not provide the results with the protocols used in the other five approaches and our approach, we re-implement their approach and report the numbers with provided official code. **Table 5** shows the comparative evaluation over all datasets, protocols, and metrics. On ImageNet-100, our

approach achieves the best results across protocols and metrics. On ImageNet-1000 and CAIFAR-100, AFC [16] obtains the best results in terms of Avg. Acc. by using special knowledge distillation. While our approach still obtains competitive results even using common knowledge distillation. With regard to forgetting on CAIFAR-100, Mnemonics [22] achieves the best results. While we achieve the better results in terms of Avg. Acc. and Acc.. From this experiment, we conclude that our approach is not only robust against distribution shifts, but also achieves favorable accuracy with minimal forgetting.

5. Conclusion

In this paper, we improve continual learning by further considering its out-of-distribution robustness, beyond just paying attention to the catastrophic forgetting problem. We do so by proposing a shape-texture debiased continual learning algorithm, which is able to learn generalizable and robust representations. With the proposed shape-texture debiased data generation and online shape-texture debiased self-distillation, standard continual learning can be easily transformed to shape-texture debiased continual learning. Our algorithm obtains favorable performance compared to current continual learning algorithms on improving generalization and reducing forgetting, especially on improving out-of-distribution robustness. The flatter local minima explained the benefits of our algorithm. Our algorithm can be easily combined with new advanced architectures like vision transformer, and applied to more challenging scenarios like exemplar-free continual learning.

References

- [1] *Benchmarking neural network robustness to common corruptions and perturbations*, 2019. 4, 5
- [2] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *ICCV*, 2021. 3
- [3] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018. 1, 2
- [4] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018. 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 *CVPR*, 2009. 4
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7
- [7] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, 2020. 1, 2, 3, 5, 8
- [8] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *CVPR*, 2022. 3, 4, 5
- [9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2, 3, 6
- [10] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1, 2
- [11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 4
- [12] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *NeurIPS*, 2020. 2, 6
- [13] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2
- [14] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 1, 2, 3, 4, 5, 8
- [15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3
- [16] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *CVPR*, 2022. 8
- [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. 1, 2
- [18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4
- [19] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS*, 2018. 6
- [20] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. In *ICLR*, 2021. 1, 2, 6
- [21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 40(12):2935–2947, 2017. 1, 2
- [22] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *CVPR*, 2020. 1, 2, 3, 4, 5, 8
- [23] 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, 2
- [24] Zichen Miao, Ze Wang, Wei Chen, and Qiang Qiu. Continual learning with filter atom swapping. In *ICLR*, 2022. 8
- [25] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *NeurIPS*, 2020. 6
- [26] Sudhanshu Mittal, Silvio Galesso, and Thomas Brox. Essentials for class incremental learning. In *CVPR*, 2021. 1, 2, 3, 5, 8
- [27] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 1, 2, 3, 5
- [28] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 1, 2
- [29] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective. In *ICML*, 2020. 2
- [30] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *ECCV*, 2020. 4, 5
- [31] Guido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. 2
- [32] Max Welling. Herding dynamical weights to learn. In *ICML*, 2009. 2
- [33] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019. 1, 2
- [34] Pei Yu, Yinpeng Chen, Ying Jin, and Zicheng Liu. Improving vision transformers for incremental learning. *arXiv preprint arXiv:2112.06103*, 2021. 7, 8
- [35] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017. 1, 2