Online Class Incremental Learning on Stochastic Blurry Task Boundary via Mask and Visual Prompt Tuning

Jun-Yeong Moon* Keon-Hee Park* Jung Uk Kim† Gyeong-Moon Park†

Kyung Hee University, Yongin, Republic of Korea

{moonjunyyy, pgh2874, ju.kim, gmpark}@khu.ac.kr

Abstract

Continual learning aims to learn a model from a continuous stream of data, but it mainly assumes a fixed number of data and tasks with clear task boundaries. However, in real-world scenarios, the number of input data and tasks is constantly changing in a statistical way, not a static way. Although recently introduced incremental learning scenarios having blurry task boundaries somewhat address the above issues, they still do not fully reflect the statistical properties of real-world situations because of the fixed ratio of disjoint and blurry samples. In this paper, we propose a new Stochastic incremental Blurry task boundary scenario, called Si-Blurry, which reflects the stochastic properties of the real-world. We find that there are two major challenges in the Si-Blurry scenario: (1) intra- and inter-task forgettings and (2) class imbalance problem. To alleviate them, we introduce Mask and Visual Prompt tuning (MVP). In MVP, to address the intra- and inter-task forgetting issues, we propose a novel instance-wise logit masking and contrastive visual prompt tuning loss. Both of them help our model discern the classes to be learned in the current batch. It results in consolidating the previous knowledge. In addition, to alleviate the class imbalance problem, we introduce a new gradient similarity-based focal loss and adaptive feature scaling to ease overfitting to the major classes and underfitting to the minor classes. Extensive experiments show that our proposed MVP significantly outperforms the existing state-of-the-art methods in our challenging Si-Blurry scenario. The code is available at `https://github.com/moonjunyyy/Si-Blurry`

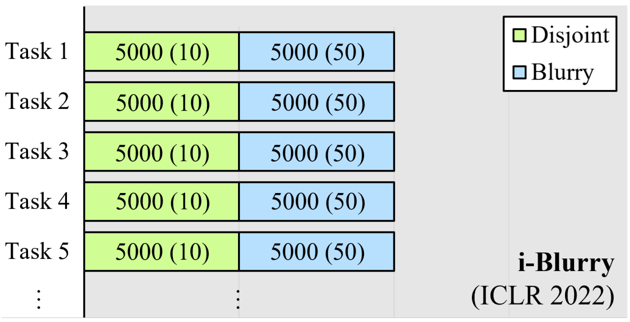
1. Introduction

Continual learning involves constantly learning from a stream of data while having limited access to previously seen information. In this scenario, unlike humans who can retain and apply their prior knowledge to new situations, modern deep neural networks face a challenge of *catastrophic forgetting* [30, 11]. To overcome this challenge, various approaches are being explored [32, 31, 13, 21, 8]. However, these traditional continual learning scenarios have clear task boundaries, where one can distinguish tasks with input data, unlike in the real-world. In real-world applications, clear task boundaries are often absent and access to data is limited to small portions at a time. This is referred to as online learning with blurry task boundaries [3].

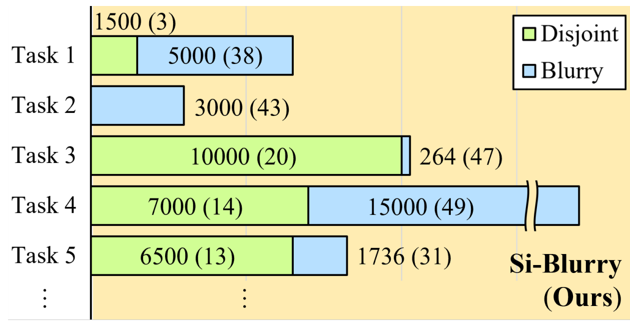
There are many cases of class emerging or disappearing like the stock market or e-commerce. To address this, the i-Blurry scenario [18] has been recently proposed, which combines disjoint continual learning and blurry task-free continual learning. Although i-Blurry somewhat alleviates the above issue, it does not fully capture the complexity of real-world data, because i-Blurry has the fixed number of classes between tasks. Figure 1(a) shows the i-Blurry scenario that contains the static number of classes in each task. In the real-world scenarios, the number of classes and tasks vary dynamically as illustrated in Figure 1(c) and 1(d). That is, as samples of a specific class continuously disappear or appear in the data stream, distribution of the data is dynamically changing.

To reflect the dynamic distribution of the real-world data, we propose a novel Stochastic incremental Blurry (Si-Blurry) scenario. We adopt a stochastic approach to imitate the chaotic nature of the real-world. As shown in Figure 1(b), our Si-Blurry scenario is capable of effectively simulating not only newly emerging or disappearing data but also irregularly changing data distribution. In the Si-Blurry scenario, we find that there are two main reasons for performance degradation: (1) intra- and inter-task forgettings, and (2) class imbalance problem. First, the continuous change of classes between batches causes intra- and inter-task forgettings, which make the model difficult to retain previously learned knowledge. Second, ignorance of minor classes

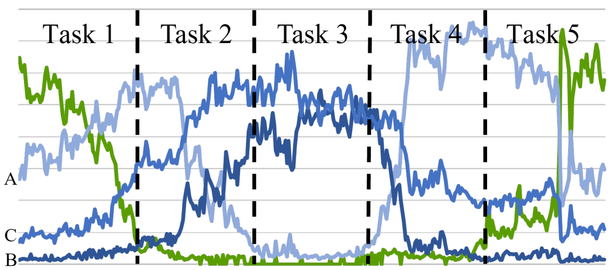
*Equally contributed
†Corresponding author



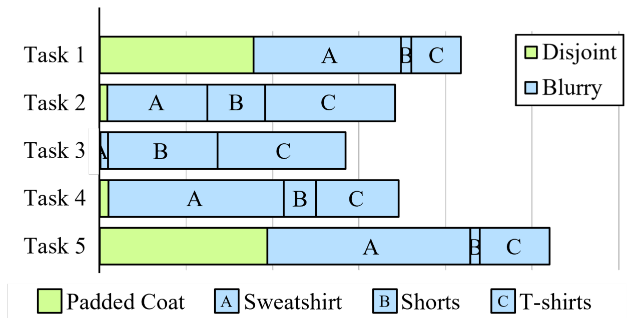
(a) Visualization of i-Blurry scenario [18].



(b) Visualization of our Si-Blurry scenario.



(c) Nomalized search history from December 28, 2021, to December 28, 2022, NAVER Search Trend API : https://www.ncloud.com/product/applicationService/searchTrend.



(d) Example of task generated from real-world data.

Figure 1. Comparision of (a) i-Blurry scenario, (b) Si-blurry scenario, and (c), (d) real-world data. (a) i-Blurry scenario consists of the static number of classes in each task. all the tasks have the same number of disjoint classes and blurry classes. (b) The Si-Blurry scenario has an everchanging number of classes. The disjoint classes and blurry classes are different with each task. It denotes the unpredictable traits of the real-world. (c) and (d) show the stream of real data. Comparing (a), (b), and (d), Si-blurry is alike real-world task configuration.

and overfitting to major classes worsen the class imbalance problem in the Si-Blurry scenario. Minor class ignorance occurs by insufficient consideration of a few samples which belong to minor classes in training, and overfitting on major classes makes the model biased to a large number of samples that belong to major classes or disjoint classes.

To deal with the aforementioned problems, we propose a novel online continual learning method called Mask and Visual Prompt tuning (MVP). We propose instance-wise logit masking and contrastive visual prompt tuning loss to alleviate intra- and inter-task forgettings by making classification easier and allowing prompts to learn the knowledge for each task effectively. Moreover, we propose a gradient similarity-based focal loss to prevent the problem of minor class ignorance. This method boosts learning of the ignored samples of minor classes in a batch, so that the samples for minor classes can be considered intensively. We also propose adaptive feature scaling to address the problem of overfitting to major classes. This method measures the marginal benefit [7] of learning from a sample and prevents our model from learning already sufficiently trained samples.

We summarize our main contributions as follows:

- We introduce a new incremental learning scenario, coined Si-Blurry, which aims to simulate a more realistic continual learning setting that the neural networks continually learn new classes online while a task boundary is stochastically varying.
- We propose an instance-wise logit masking and contrastive visual prompt tuning loss to prevent the model from intra-task and inter-task forgettings.
- To solve the class imbalance problem, we propose a new gradient similarity-based focal loss and adaptive feature scaling for minor-class ignorance and overfitting on major classes.
- We experimentally achieved significantly high performance compared to existing methods, supporting that our proposed method shows overwhelming performance and solves the problems of Si-Blurry in CIFAR-100, Tiny-ImageNet, and ImageNet-R.

2. Related Work

2.1. Disjoint Continual Learning

Class Incremental Learning (CIL) [12] assumes that each task contains distinct classes not overlapping with another and that the class observed once in a task never appears again in subsequent tasks. CIL is categorized into the (1) regularization-based method, (2) replay-based method, (3) parameter isolation method, and (4) prompt-based method. The regularization-based method uses previous knowledge for regularizing the network while training new tasks [23, 17, 4, 27, 42]. The replay-based method stores a few samples from the old task and replays them in the new task to mitigate catastrophic forgetting [34, 38, 2, 6, 29]. The parameter isolation method expands the network or consists of sub-networks in a single network for each task [37, 43, 1, 36, 44]. The prompt-based method proposed in natural language processing (NLP) for transfer learning attaches a set of learnable parameters, named prompt, to the frozen pre-trained model [40, 39, 15].

2.2. Blurry Continual Learning

Blurry Continual Learning [33, 3] assumes no new classes appear after the first task even though classes overlap across the tasks. A blurry setup has some requirements. First of all, each task is streamed sequentially. Second, the major class of each task is different. Last, a model can leverage only a small portion of data from the previous task. A blurry setup seems realistic. However, a blurry scenario has a shortage to apply real-world scenarios in that observing new classes is commonplace in a real-world scenario. i-Blurry [18] proposes a more realistic setting that considers a blurry scenario with a class-incremental setting. However, the i-Blurry scenario also has a limitation of properly reflecting the real-world scenario due to: (1) the same number of new classes appearing in every task, and (2) new classes and blurry classes having the same proportion in every task. This is why we propose a stochastic incremental blurry scenario, which focuses on the stochastic property critical to real-world scenarios.

2.3. Class Imbalance in Continual Learning

Class imbalance, known as the long-tail problem, is that the classes are not represented equally in the classification task. Class imbalance is common in the real-world and it can cause inaccurate prediction performance in classification problems. In continual learning, the replay-based method suffers severe catastrophic forgetting due to the inequality between stored old samples and streamed new samples [42]. To address this problem, existing methods consider gradient information to get the knowledge of prior tasks during training [29, 2], episodic memory management to enhance model performance by sampling effective samples [5, 28, 41], and calibration of the bias [42]. However, in a blurry scenario, incoming samples per class are different and this causes a class imbalance problem. Class imbalance in a task leads to bias for disjoint classes and major classes which exacerbates the training of the minority classes.

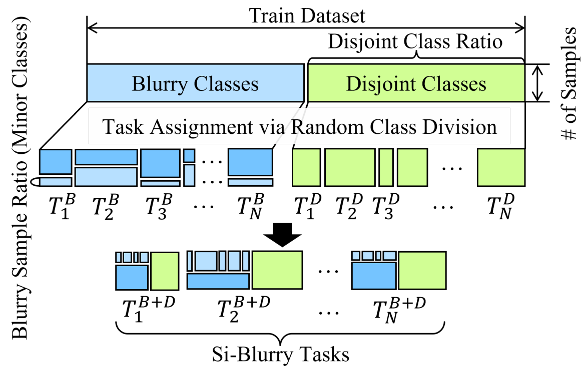


Figure 2. A configuration example of the proposed Si-Blurry scenario.

3. Stochastic Incremental Blurry Scenario

3.1. Scenario Configuration

In a real-world scenario, the quantity of input data and tasks tends to change a stochastic manner. To simulate this, we propose the Stochastic incremental-Blurry (Si-Blurry) scenario. From [18], we divide the classes into two categories using disjoint class ratio: disjoint classes and blurry classes. As shown in Figure 2, we randomly assign each blurry class and disjoint class to each blurry task (T^B) and disjoint task (T^D) by the disjoint class ratio. In blurry tasks, we gather the sample of blurry sample ratio and randomly distribute it to each task. This makes the classes on each task overlap, which blur explicit task boundary. Each task with a stochastic blurry task boundary T^{B+D} consists of T^B and T^D. Figure 1(b) shows an example of the distribution of Si-Blurry. Because the Si-Blurry task is stochastic, the batches get more diverse and imbalanced. As a result, there are lack of explicit task boundaries and the data imbalance, which pose significant challenges to formal continual learning methods. We define two problems that are exacerbated on Si-Blurry in the following subsections.

3.2. Intra- and Inter-Task Forgettings

First, intra-task forgetting can be interpreted as inter-batch forgetting. This problem also presents in joint training but is largely addressed by randomizing the sample order in each batch. However, in online learning scenarios, each sample is only presented once, making it impossible to apply this strategy. Additionally, the stochastic nature of Si-Blurry creates more diverse batches, thereby intensifying the issue of intra-task forgetting. Intra-task forgetting can severely limit the ability of the model to learn and gen-

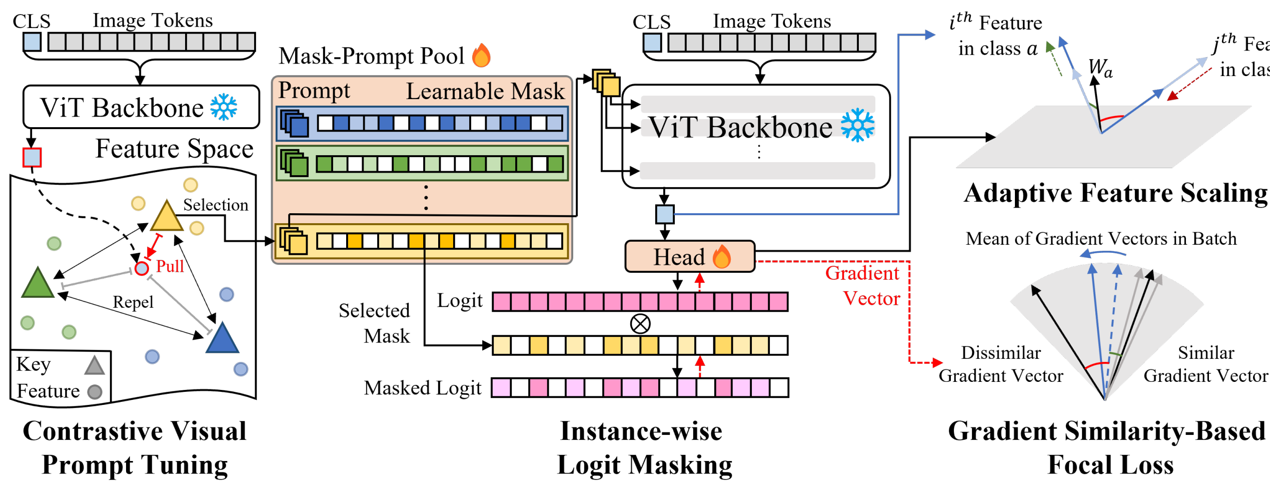


Figure 3. Illustration of the process of the proposed method. Firstly, a pre-trained ViT is used to extract features, and a key is compared with the feature by cosine distance to select the matched prompt and learnable mask. We use the prompt tuning [39] method with deep prompting [15]. And the selected mask is multiplied element-wise to logit. To train the key, we use a contrastive manner and weighted key to prevent key grouping. After the forward pass, the weight vector of the label class in the classification head and feature are compared to measure the marginal benefit of the sample. In the backward path, we compare the mean of the gradient vector in the batch and the gradient vector of each sample to measure the ignored sample. Finally, by scaling the feature and the cross-entropy loss, we effectively handle the forgetting and class imbalance problems.

eralize well in the face of evolving and dynamic data distributions.

Inter-task forgetting, which refers to the phenomenon of losing previously learned knowledge of a task due to the changes in class distribution, is a major challenge in online continual learning. Si-Blurry, a scenario proposed to simulate the complexity of real-world data in a stochastic manner, does not have clear task boundaries like those in conventional continuous learning, making it difficult to handle this problem. The stochastic change in class distribution in Si-Blurry exacerbates this problem. Effective strategies need to be devised to enable the model to learn the knowledge continually from data with varying class distributions, without losing previously acquired knowledge, and any catastrophic forgetting.

3.3. Class Imbalance

Minor-class ignorance and overfitting to major classes cause the class imbalance problem. The issue of minor-class ignorance arises when the number of samples in a batch varies and this issue causes an imbalanced weighted loss. The loss imbalance leads to the negligence of the minority of samples in a batch, resulting in their poor representation while training the model. Overfitting to major classes is, conversely, a phenomenon where a class with a large number of samples deteriorates the model performance by acquiring unnecessary knowledge relating to generalization performance.

These are particularly important issues in the context of Si-Blurry, where no explicit task boundaries exist, and continual learning requires a model to capture the knowledge from a wide range of samples. Finding a solution to this problem is essential for ensuring that the model remains effective in recognizing and classifying all samples, regardless of the number of samples per class, and that it continues to learn and improve over time.

4. Mask and Visual Prompt Tuning (MVP)

4.1. Preliminary and Problem Formulation

The Si-Blurry scenario considers learning a model with only a few samples because it cannot access the whole current training data. When given the accessible data $\mathcal{B} = \{\mathbf{x_i}, y_i\}_{i=1}^{N}$, where $\mathbf{x_i} \in \mathcal{X}, y_i \in \mathcal{Y}$, we reshape the samples to a flattened patch shape $\mathbb{R}^{L\times(S^2\times C)}$ to feed into the pre-trained model $f : \mathbb{R}^{L\times(\mathrm{S}^2\times\mathrm{C})} \rightarrow \mathbb{R}^{L\times\mathrm{D}}$ which is frozen, where L, S, C, and D represent the token length, patch size, channel, and embedding dimension, respectively. A linear classifier $W \in \mathbb{R}^{D\times|\mathcal{Y}|}$ is trainable.

Previous studies demonstrate the effectiveness of using knowledge from the pre-trained model and tuning the small size of parameters [40, 39] for continual learning. To this end, we adopt the prompt tuning method [22] for online continual learning. Similarly to DualPrompt [39], we utilize a pre-trained Vision Transformer (ViT) as a feature extractor for the query. We match the query with the key to apply the contrastive visual prompt tuning loss and select the prompt.

4.2. Instance-wise Logit Masking

Existing prompt methods assume the explicit task boundary and require information on the task boundary for training, which is not feasible in Si-Blurry, where no explicit task boundary exists. The cross-entropy loss is highly effective in optimizing classification models, but it requires a proper comparison target to acquire sufficient knowledge. Unlike traditional joint training, the absence of comparison constantly leads to forgetting in online continual learning. To address this problem causing intra- and inter-task forgettings, we propose a new instance-wise logit masking technique.

To complement the prompt-based continual learning approach and further enhance the performance of the model, we introduce a learnable mask paired with prompts that helps the model to learn more intra-relevant and easier learning goals. Since the key-value mechanism is used to select each prompt, where a feature extracted by the pre-trained model serves as a query, each prompt can be responsible for a certain region of the feature space in which classes have similar extracted features. As illustrated in Figure 3, we apply the mask to logit using an element-wise product and train the mask, and then calculate cross-entropy loss which makes the mask divide the tasks into easier classification tasks. The logit masking provides the model with a scaled gradient during back-propagation, which protects the knowledge that has been sufficiently trained and encourages the learning of classes to be learned.

4.3. Contrastive Visual Prompt Tuning Loss

The logit mask assumes that each key enables the prompts to learn similar knowledge. We empirically find that the existing prompt-based method [40] converges to a single point, which renders the query selection mechanism inaccurate and meaningless. Moreover, the keys are updated continuously, which causes forgetting. To overcome these challenges and leverage the benefits of prompt-based continual learning, we propose a novel loss function, called Contrastive Visual Prompt Tuning Loss. We formulate this loss term as follows:

$$s_p = \sum_{p=1}^{P}\sum_{q=1}^{B} \exp\left(\delta\left(\mathbf{k}_p, \mathbf{q}_q\right)/(\mathcal{C}_p+1)\right),$$
$$s_n = \sum_{p=1}^{P}\sum_{q=1}^{P} \exp\left(\delta\left(\mathbf{k}_p, \mathbf{k}_q\right)/(\mathcal{C}_p+1)\right),$$
$$\mathcal{L}_{CVPT} = -\log\frac{s_n}{s_p+s_n}, \tag{1}$$

where δ is cosine distance, P denotes the size of the prompt pool, $\mathbf{q}_n \in \mathbb{R}^D$ indicates the query feature, $\mathbf{k}_n \in \mathbb{R}^D$ denotes the key of n^{th} prompt, and $\mathcal{C}_n$ means the count of selection of n^{th} prompt. In $\mathcal{L}_{CVPT}$, $(\mathcal{C}_p+1)$ plays a role of the temperature to control the softness. If $\mathcal{C}_n$ is large, the effect of loss to key lessens. As illustrated in Figure 3, the $\mathcal{L}_{CVPT}$ increases the distances between keys. Also, as the prompt learns, the keys become heavier to ensure consistency in key selection. The instance-wise logit masking coupled with $\mathcal{L}_{CVPT}$ can prevent inter-task and intra-task forgetting by ensuring that each prompt divides its responsible region and preserves the knowledge.

4.4. Gradient Similarity-based Focal Loss

Since the blurry setup has a task that comprises of imbalanced classes, it is challenging for the model to extract the knowledge of all the observed classes in the blurry setup. Due to the stochastic nature of Si-Blurry, we cannot guarantee a minimum number of samples for minor classes. To mitigate the aforementioned minor class ignorance, we propose a Gradient Similarity-based Focal loss $\mathcal{L}_{GSF}$ (GSF loss). It focuses on the loss from ignored samples leveraging ignore scores Score^{ign}. The ignore score Score_i^{ign} denotes how much a sample $\mathbf{x_i}$ is ignored by other samples during training. We use cosine distance in between a gradient vector from each sample $\nabla W_{y_i}(f(\mathbf{x_i})) \in \mathbb{R}^D$ and the averaged gradient vector $\nabla W_{y_i}(f(\mathcal{B})) \in \mathbb{R}^D$ from accessible data $\mathcal{B}$ to yield an ignore score for a sample $\mathbf{x}_i$. We formulate ignore score and GSF loss as:

$$\nabla W_{y_i}(f(\mathcal{B})) = \frac{1}{|\mathcal{B}|}\sum_{(\mathbf{x},y)\in\mathcal{B}} \nabla W_{y_i}(f(\mathbf{x})),$$
$$\text{Score}_i^{ign} = \delta\left(\nabla W_{y_i}\left(f\left(\mathbf{x}_i\right)\right), \nabla W_{y_i}(f(\mathcal{B}))\right), \tag{2}$$
$$\mathcal{L}_{GSF} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\text{Score}_i^{ign}\right)^{\gamma}\cdot\mathcal{L}_{CE}\left(\hat{y}_i, y_i\right), \tag{3}$$

where $(\mathbf{x}, y) \in \mathcal{B}$ is a training sample, $\mathcal{L}_{CE}$ is a cross-entropy loss, and W_{y_i} is the weights of corresponding label, respectively. Eq. 2 represents ignore score for a sample $\mathbf{x}_i$. When the Score^{ign} has a high value, it implies the model is hard to extract the knowledge from the sample, whereas the low Score^{ign} means vice versa. Leveraging Score^{ign}, we can emphasize the loss from the ignored sample and capture more knowledge of minor classes than before as illustrated in Figure 3. Eq. 3 represents GSF loss which considers the amount of ignorance. In [24], the focal loss dynamically scales cross-entropy loss considering confidence in the correct class. Our proposed GSF loss also dynamically scales cross-entropy loss. However, our loss scales the cross-entropy loss considering ignore score. Ignore score is the degree of ignorance that is conceptually different from confidence. GSF loss can mitigate the class ignorance problem which minor classes highly suffered, and enables balanced class learning.

4.5. Adaptive Feature Scaling

In online learning, the model cannot access all the data of the current task but access a few samples. In our novel Si-Blurry scenario, each task has a class imbalance problem. When the model access training data, there are no or few samples of minor classes. It makes the model overfit to major classes and newly streamed disjoint classes. To mitigate the overfitting problem caused by the class imbalance, we propose Adaptive Feature Scaling (AFS) which expands or contracts the feature vector considering the marginal benefit score Score^{MB}. Using the Score^{MB}, the model can learn new knowledge from the accessible data while preserving the knowledge from inaccessible prior data.

As prior works [9, 25, 26] suggest, the similarity between the feature vector and the weights from the last fully connected layer relates to the prediction when the model is trained by cross-entropy loss with softmax function. We propose a marginal benefit score Score^{MB} that represents how similar the feature vector is with the weights of the corresponding label. Leveraging this, we estimate the marginal benefit from the given instance and adjust the model updates by the given instance. We can calculate Score^{MB} as follows:

$$\mathrm{Score}_i^{MB} = \delta\left(f\left(\mathbf{x}_i\right), W_{y_i}\right) + m, \tag{4}$$

$$\mathbf{h_i} = \frac{f(\mathbf{x}_i)}{\mathrm{Score}_i^{MB}}, \tag{5}$$

$$\widehat{y}_i = W(\mathbf{h_i}) \tag{6}$$

where m is a margin, and $\mathbf{h_i}$ is a scaled feature vector by Score_i^{MB}. When the Score^{MB} has a high value, it implies the given sample has a large marginal benefit. The Score^{MB} reduces the feature vector to increase the expected loss value. Enlarged expected loss makes the model learn enough knowledge from the given sample. In contrast, when the Score^{MB} has a low value, it implies the given sample has a little marginal benefit for the model. In this case, a feature vector is expanded to decrease the expected loss value. The model is less trained due to the curtailed expected loss. We estimate the marginal benefit that can be extracted from a sample and scale up and down the feature vector considering the marginal benefit.

To overcome the class imbalance problem, we propose two components that seem similar: gradient similarity-based focal loss (GSF) and adaptive feature scaling (AFS). Although GSF and AFS look similar, their main roles are different. GSF emphasizes learning minor classes to tackle the class imbalance problem in the task. AFS regularizes learning major classes to address the overfitting problem.

Finally, we train our model in an end-to-end manner. The total loss for our method is defined as:

$$\mathcal{L}_{\mathrm{total}} = (1-\alpha)\mathcal{L}_{\mathrm{CE}} + \alpha\mathcal{L}_{GSF} + \mathcal{L}_{CVPT}\,, \tag{7}$$

where $\mathcal{L}_{\mathrm{CE}}$ is cross entropy loss with instance-wise logit masking. We use hyperparameter α for the balanced training and γ at the gradient similarity-based focal loss to scale the ignore score.

5. Experiments

5.1. Experimental Details

Datasets. We conducted experiments on three datasets: CIFAR-100 [19], Tiny-ImageNet [20], and Imagenet-R [14] with 60,000, 100,000, and 30,000 samples, respectively, and 100, 200, and 200 classes. For the Si-Blurry scenario, we set a disjoint class ratio to 50% and a blurry sample ratio to 10%. We evaluated each method on five independent seeds to empirically search the difficulty of Si-Blurry.

Baselines. We compared our method with the state-of-the-art methods, including replay-based methods such as ER [35], Rainbow Memory (RM) [3], and CLIB [18], regularization-based method LwF [23], replay-regularization combined method EWC++ [17], and prompt-based methods L2P [40] and DualPrompt [39]. We used finetuning and linear probing as lower-bound of our scenario. We implemented all methods with the fixed pre-trained Vision Transformer [10], but the EWC++ and finetuning conducted full finetuning.

Implementation Details. We used Adam optimizer [16] to train our MVP with 0.005 learning rate. For the existing methods [35, 3, 18, 23, 17, 40, 39], we re-implemented them by following the setting (*e.g.,* learning rate) of original papers in order to compare the performance with our MVP in the Si-Blurry scenario. We set the batch size to 32. For training our MVP, we set α to 0.5 and γ to 2.0. We empirically observed that our method is robust to α and γ. Please note that although our MVP is designed to be memory-free, it also can use a memory buffer if it is needed. To this end, we adopt the replay buffers capable of storing 500 and 2,000 samples. We call our MVP with replay buffers "MVP-R".

Evaluation Metrics. To evaluate the online learning performance, we used two metrics, A_{AUC} and A_{Last}. A_{AUC} metric, proposed in [18], measures the performance of anytime inference. Anytime inference assumes that inference queries can occur anytime during training with exposed classes. A_{Last} measures the inference performance after the train is ended. In the real-world, the model needs to offer the proper services whenever a client is needed. In this point of view, Evaluating performance by A_{Last} and A_{AUC} is appropriate to benchmark the performance in the real-world.

Buffer Size	Method	CIFAR-100		Tiny-ImageNet		ImageNet-R	
		A_{AUC}	A_{Last}	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}
0	Finetuning	19.71±3.39	10.42±4.92	15.50±0.74	10.42±4.92	7.51±3.94	2.29±0.85
	Linear Probing	49.69±6.09	23.07±7.33	42.15±2.79	21.97±6.43	29.24±1.26	16.87±3.14
	LwF [23]	55.51±3.49	36.53±10.96	49.00±1.52	27.47±7.59	31.61±1.53	20.62±3.67
	L2P [40]	57.08±4.43	41.63±12.73	52.09±1.92	35.05±5.73	29.65±1.63	19.55±4.78
	DualPrompt [39]	67.07±4.16	56.82±3.49	66.09±2.00	48.72±3.41	40.11±1.27	29.24±4.63
	MVP (Ours)	**68.10±4.91**	**62.59±2.38**	**68.95±1.33**	**52.78±2.08**	**40.60±1.21**	**31.96±3.07**
500	ER [35]	65.57±4.77	60.68±1.15	59.46±1.81	40.60±2.71	40.31±1.33	28.85±1.43
	EWC++ [17]	34.54±5.19	25.62±3.35	55.05±1.75	34.88±3.65	18.62±1.00	11.36±2.40
	RM [3]	40.86±3.32	23.94±0.61	31.96±0.80	7.43±0.27	18.31±1.09	4.14±0.18
	CLIB [18]	69.68±2.20	67.16±0.72	60.11±1.53	48.97±1.48	37.18±1.52	29.51±0.98
	MVP-R (Ours)	**76.06±4.22**	**79.32±1.28**	**76.52±0.73**	**65.19±0.58**	**49.07±1.47**	**44.17±1.72**
2,000	ER [35]	69.86±4.08	71.81±0.69	66.75±1.13	55.07±1.28	45.74±1.35	38.13±0.32
	EWC++ [17]	47.75±5.35	46.93±1.44	64.92±1.21	53.04±1.53	30.20±1.31	21.28±1.88
	RM [3]	53.27±3.00	65.51±0.55	47.26±1.13	44.55±0.37	27.88±1.29	24.25±0.99
	CLIB [18]	71.53±2.61	72.09±0.49	65.47±0.76	56.87±0.54	42.69±1.30	35.43±0.38
	MVP-R (Ours)	**78.65±3.59**	**84.42±0.44**	**80.67±0.75**	**74.34±0.32**	**52.47±1.45**	**50.54±2.08**

Table 1. Average accuracy of continual learning methods on Si-Blurry scenario. For the comparison, we adopt the CIFAR-100, Tiny-ImageNet, and ImageNet-R datasets. Note that MVP-R indicates our MVP with a memory buffer.

Method	Components		Memory = 0		Memory = 2,000	
	Mask	Cont	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}
Baseline	-	-	67.07±4.16	56.82±3.49	75.26±5.02	80.72±0.83
MVP (Ours)	✓		67.64±3.81	58.52±3.28	77.83±0.35	84.26±0.04
		✓	66.37±4.59	58.63±1.18	76.67±1.98	83.32±0.40
	✓	✓	**68.08±6.46**	**60.20±3.28**	**77.85±0.04**	**84.28±0.15**

Method	Components		Memory = 0		Memory = 2,000	
	GSF	AFS	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}
Baseline	-	-	67.07±4.16	56.82±3.49	75.26±5.02	80.72±0.83
MVP (Ours)	✓		67.45±5.21	56.11±3.20	77.34±2.16	83.75±0.53
		✓	67.45±3.78	57.93±2.11	77.86±2.09	84.31±0.20
	✓	✓	**67.66±3.47**	**58.28±2.95**	**78.28±3.67**	**84.41±0.21**

Table 2. Ablation studies of instance-wise logit masking and the contrastive visual prompt tuning for alleviating inter and intra-task forgetting. 'Mask' and 'Cont' represent instance-wise logit masking and contrastive visual prompt tuning, respectively. The lower table represents the ablation study of gradient similarity-based focal loss and adaptive feature scaling. GSF and AFS denote gradient similarity-based focal loss and adaptive feature scaling, respectively.

5.2. Results on the Si-Blurry Scenario

We compared our MVP to other online CL methods in the Si-Blurry scenario. The results are shown in Table 1. In the Si-Blurry scenario, our MVP method outperformed all other comparison methods. On CIFAR-100, our method without a memory buffer outperformed EWC++ and RM, where they used a memory buffer size of 500. Furthermore, our method with a memory buffer size of 500 overwhelmed other methods with large margins. On Tiny-ImageNet, our MVP also outperformed other methods. Note that the RM Score was especially low on Tiny-ImageNet. We believe that the RM method trains the model highly focusing on samples stored in a memory buffer and Tiny-ImageNet contains 100,000 training images. The limited memory buffer could not cover the vast number of samples. So RM method suffered from overfitting. On ImageNet-R, all comparison methods scored low performance whether a memory buffer was used or not. Nonetheless, MVP showed a standing-out performance score. The score gap between MVP and other methods got bigger when the memory buffer was used. Our

Method	Components				Memory = 0		Memory = 2,000	
	Mask	Cont	GSF	AFS	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}
Baseline	-	-	-	-	67.07±4.16	56.82±3.49	75.26±5.02	80.72±0.53
MVP (Ours)	✓	✓			68.08±6.46	60.20±3.28	77.85±0.04	84.28±0.15
			✓	✓	67.66±3.47	58.28±2.95	78.28±3.67	84.41±0.21
	✓	✓	✓	✓	**68.10±4.91**	**62.59±2.38**	**78.65±3.59**	**84.42±0.44**

Table 3. Ablation experiment coupled with the best condition for inter and intra-task forgetting and class imbalance. 'Mask' and 'Cont' represent instance-wise logit masking and contrastive visual prompt tuning and GSF and AFS denote gradient similarity-based focal loss and adaptive feature scaling, respectively.

method MVP got an outperforming score without memory and performed much better when memory is used.

5.3. Ablation Study

Table 2 shows an ablation study on CIFAR-100 for the proposed components in our method. We set the ablation study whether using a memory buffer or not. We set the Dual Prompt as a baseline for both cases. The ablation study showed that each component of our method was beneficial to the performance.

As shown in the upper table from Table 2, instance-wise logit masking and contrastive visual prompt tuning loss scored a better performance without memory than the performance of the baseline in both A_{AUC} and A_{last}. Moreover, using both of them scored better performance than the performance of each component. It implies instance-wise logit masking and contrastive visual prompt tuning loss makes a complementary performance. Instance-wise logit masking helps to enhance accuracy by making the task easy by masking irrelevant classes. So, the model can learn representative knowledge more efficiently.

We also investigated the GSF loss and AFS in the lower table from Table 2. Using both GSF and ASF scored the highest performance among the performance of the lower table. It showed that Each of GSF and AFS can alleviate class imbalance. In addition, using both of them was more helpful to class imbalance considerably. The lower table implies that class imbalance is a crucial problem for prediction performance in Si-Blurry.

In Table 3, We compared the set of components for the inter and intra-task forgettings (instance-wise logit masking and contrastive visual prompt tuning loss) with the set of components for the class imbalance (GSF and AFS). When we used all the components, we observed the performance improvement from 67.07±4.16 to 68.10±4.91 in A_{AUC} and from 56.82±3.49 to 62.59±2.38 in A_{Last}. This showed that our methods are synergistic.

5.4. Comparison between i-Blurry and Si-Blurry

The comparative analysis, as illustrated in Table 4, encompasses a comprehensive evaluation of the best and worst performance metrics for both existing methods and our proposed approach across two distinct scenarios: i-Blurry [18] and Si-Blurry. In the context of the i-Blurry scenario, the results underscore the superiority of our proposed MVP method, outperforming other existing methodologies in both the best and worst case scenarios. This demonstrates the consistency and robustness of our method in i-Blurry settings. Notably, our method even surpasses the state-of-the-art CLIB method, establishing its prowess in challenging real-world scenarios.

Furthermore, the insights drawn from our evaluation extend beyond the i-Blurry scenario. Our approach yields improved performance not only in the i-Blurry context but also in the Si-Blurry scenario. It is worth highlighting that while some existing methods exhibited a decline in performance within the Si-Blurry scenario, it is indicative of the increased complexity and realism inherent in Si-Blurry, rendering it a more intricate challenge than the i-Blurry scenario. The robust performance in both i-Blurry and Si-Blurry scenarios underscores the adaptability and effectiveness of our method in tackling varying levels of complexity and realism.

5.5. Disjoint Sample Ratio

We tested our method at various disjoint class ratios. We compared the performance of the model in the extreme cases of disjoint only and no disjoint, following experiment of [18]. Table 5 demonstrated that the MVP method maintained a robust and high performance across various disjoint class ratios. This suggested that our method is capable of achieving excellent results in both existing blurry and class incremental learning scenarios.

5.6. Blurry Sample Ratio

We tested our method at various blurry sample ratios, following the experiment of [18]. Table 6 refers that the MVP method maintained robust high performance while the blurry sample increased to half. Considering these results with the results from Table 6, we could conclude that our method has the ability to improve performance in a variety of data distributions that we might encounter in the real

Case	i-Blurry			Si-Blurry		
	CLIB [18]	DP[39]	MVP-R (2,000)	CLIB [18]	DP [39]	MVP-R (2,000)
Best case	72.56	67.51	84.69	72.91	61.68	84.89
Worst case	71.86	62.57	83.68	71.78	53.68	83.80
Average	72.12±0.38	64.90±1.96	**84.44±0.43**	72.09±0.49	56.82±3.49	**84.42±0.44**

Table 4. The analysis of continual learning methods between the i-Blurry scenario and our Si-Blurry scenario on the CIFAR-100 [19] dataset. 'MVP-R (2,000)' indicates our method with a memory buffer size of 2,000. The worst and best case denote the lowest and highest score respectively among the 5 independent runs, and the average score is the A_{Last} score.

Disjoint Class Ratio	0		50		100	
	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}
DualPrompt [39]	**68.85±2.77**	72.31±9.19	67.07±4.16	56.82±3.49	71.45±1.67	48.68±3.47
MVP (Ours)	67.86±2.62	**73.83±8.34**	**68.10±4.91**	**62.59±2.38**	**73.35±2.63**	**53.40±5.49**

Table 5. We compared the performance of MVP and DualPrompt at different disjoint class ratios. Experimental results show that MVP can be robust to changing disjoint class ratios.

Blurry Sample Ratio	10		30		50	
	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}	A_{AUC}	A_{Last}
DualPrompt [39]	67.07±4.16	56.82±3.49	70.58±2.05	59.47±7.38	68.08±5.56	49.93±2.82
MVP (Ours)	**68.10±4.91**	**62.59±2.38**	**71.10±2.10**	**63.02±6.68**	**70.58±2.05**	**59.47±7.38**

Table 6. We compared the performance of MVP and DualPrompt at different blurry sample ratios. Experimental results show that MVP can be more robust to a number of blurry samples than the existing method.

world. This showed that the MVP method has the ability to effectively solve real-world problems.

6. Conclusion

In this paper, we found that the previous CL scenarios fall short in their efforts to reflect reality. We designed a new scenario, Si-blurry, to simulate the complexity of a real-world data stream. In the Si-Blurry scenario, we could find two main branches of problems that degrade online learning performance: intra-task and inter-task forgetting, and class imbalance. To mitigate these problems in the Si-Blurry scenario, we proposed a Mask and Visual Prompt tuning (MVP) consisting of instance-wise logit masking, contrastive visual prompt tuning loss, gradient similarity-based focal loss, and adaptive feature scaling. Our method in the Si-Blurry scenario outperforms the existing CL methods. We believe that the Si-Blurry scenario is a step-forward scenario that reflects the real-world scenario.

Although our MVP shows a standing-out performance, it has some limitations. Different from prior work [40], we selected only one prompt from the pool to facilitate mask learning and prevent the key converging problem. However, selecting multiple keys can achieve knowledge sharing and reduce the risk of missed selections. It suggests further work to train instance-wise masks taking these advantages. Also, batch-wise calculation of the ignorance score makes the method sensitive to batch size. Batch-agnostic online learning is suggested as another further work.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1G1A1094379), and in part by MSIT (Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2023-RS-2023-00258649) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068, and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)).

References

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

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual

learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[3] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *CVPR*, 2021. 1, 3, 6, 7

[4] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018. 3

[5] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018. 3

[6] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. 3

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

[8] 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. *TPAMI*, 2021. 1

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 6

[10] 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. *arXiv preprint arXiv:2010.11929*, 2020. 6

[11] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. 1

[12] Alexander Gepperth and Barbara Hammer. Incremental learning algorithms and applications. In *ESANN*, 2016. 3

[13] Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, and Hinrich Schuetze. Neural topic modeling with continual lifelong learning. In *ICML*, 2020. 1

[14] 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. 6

[15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 3, 4

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[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. 3, 6, 7

[18] Hyunseo Koh, Dahyun Kim, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on class incremental blurry task configuration with anytime inference. In *International Conference on Learning Representations*. 1, 2, 3, 6, 7, 8, 9

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. 6, 9

[20] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 6

[21] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *IF*, 2020. 1

[22] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 4

[23] Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 2017. 3, 6, 7

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 6

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016. 6

[27] Yu Liu, Sarah Parisot, Gregory Slabaugh, Xu Jia, Ales Leonardis, and Tinne Tuytelaars. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *ECCV*, 2020. 3

[28] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *CVPR*, 2020. 3

[29] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. volume 30, 2017. 3

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

[31] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *NN*, 2019. 1

[32] B. Pfülb and A. Gepperth. A comprehensive, application-oriented study of catastrophic forgetting in DNNs. In *ICLR*, 2019. 1

[33] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV*, 2020. 3

[34] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 3

[35] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. 2019. 6, 7

[36] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 3

[37] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. 2017. 3

[38] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *ECCV*, 2022. 3

[39] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, 2022. 3, 4, 6, 7, 9

[40] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, 2022. 3, 4, 5, 6, 7, 9

[41] Max Welling. Herding dynamical weights to learn. In *ICMR*, 2009. 3

[42] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019. 3

[43] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *CVPR*, 2021. 3

[44] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017. 3