Computing and Informatics, Vol. 44, 2025, 1202–1228, doi: 10.31577/cai_2025_5_1202

ADOCIL: ENHANCING IMAGE CLASSIFICATION WITH ATTENTION DISTILLATION FOR ONLINE CLASS-INCREMENTAL LEARNING

Jinyong Cheng, Mengyun Chen, Baoyu Du, Min Guo*

Key Laboratory of Computing Power Network and Information Security
Ministry of Education, Shandong Computer Science Center
(National Supercomputer Center in Jinan)
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, China
&
Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing
Shandong Fundamental Research Center for Computer Science
Jinan, China
e-mail: chenmengyun2024@163.com, guomin@qlu.edu.cn

Abstract. Catastrophic forgetting is a major challenge for online class-incremental learning. Existing replay-based methods have achieved a certain degree of effectiveness, but are limited by not considering the quality of the samples and the key semantic information in a single-pass data stream. To address these issues, we proposed the framework of Online Class-Incremental Learning Based on Attention Distillation(ADOCIL), which consists of three parts. A two-stage sampling method is used in the replay stage to improve the quality of the samples taken. Meanwhile, we introduced the Attention-based Dual-View Consistency (ADVC), which enables the model to fully explore the critical semantic information within a single-pass data stream. In addition, to further mitigate the problem of catastrophic forgetting, we introduced attention distillation to map the attentional map of the teacher model to the student model, thus solving the problem of forgetting historical tasks. Extensive experiments demonstrated the effectiveness of ADOCIL.

Keywords: Catastrophic forgetting, class-incremental learning, two-stage sampling, ADVC, attention distillation

* Corresponding author

1 INTRODUCTION

In recent years, deep learning in the field of computer vision has witnessed significant development, with deep neural networks achieving remarkable results in various domains [1, 2, 3]. However, real-world open environments often generate data in a streaming fashion [4, 5]. To enable deep learning models to continuously accept and learn new knowledge, class-incremental learning has emerged. Class-incremental learning aims to address a significant problem that occurs when models learn new classes – catastrophic forgetting [6, 7, 8, 9]. Catastrophic forgetting refers to the phenomenon where a model may forget previously learned knowledge when learning new categories or tasks, leading to poor performance on tasks that it was once proficient in. For example, in Figure 1, we depict the setup of class-incremental learning, where each new task includes information about old classes. Our goal is to enable the model to correctly classify both new and old classes. For instance, in Task 2, we expect the model to accurately recognize the new classes "lion" and "sheep", while still being able to identify the old classes "cat" and "dog". However, the issue of catastrophic forgetting often results in a significant drop in the recognition accuracy of the old classes "cat" and "dog". This problem hinders the model's ability to continually accumulate knowledge, limiting the application of deep learning in open environments. Therefore, effectively fighting catastrophic forgetting becomes a core challenge for class-incremental learning [10, 11, 12].

To address the issue of catastrophic forgetting, researchers have proposed various methods [13, 14, 15], including techniques based on replay, parameter isolation, and regularization. Among these, replay-based methods can be further categorized into memory-based and generate replay methods. While regularization and parameter isolation methods can fight catastrophic forgetting, their effectiveness remains limited compared to replay-based methods. Replay-based methods store a portion of previously trained old data in a buffer and, during the learning of new tasks, select relevant old data from the buffer for replay. In this process, the selection of valid old data is crucial for success. However, existing replay-based methods tend to ignore situations where samples have quality issues when selecting samples for replay.

We have recognized that real-world images often suffer from issues like blurriness, shadows, and low clarity [16, 17]. Directly selecting such samples for training would severely impact the model's performance. Therefore, we proposed Two-Stage Sampling to alleviate this problem. In the first stage, we calculate the average gradient of images in the storage buffer and sort them in descending order based on their average gradient values. This allows us to prioritize images with higher average gradients, as they typically contain more details and important information. In the second stage, we further select those images that cause the largest gradient changes to the new samples based on the images selected in the first stage. These images will have a greater impact on training as they retain key information about the old class, benefiting the updated network. By adopting this strat-

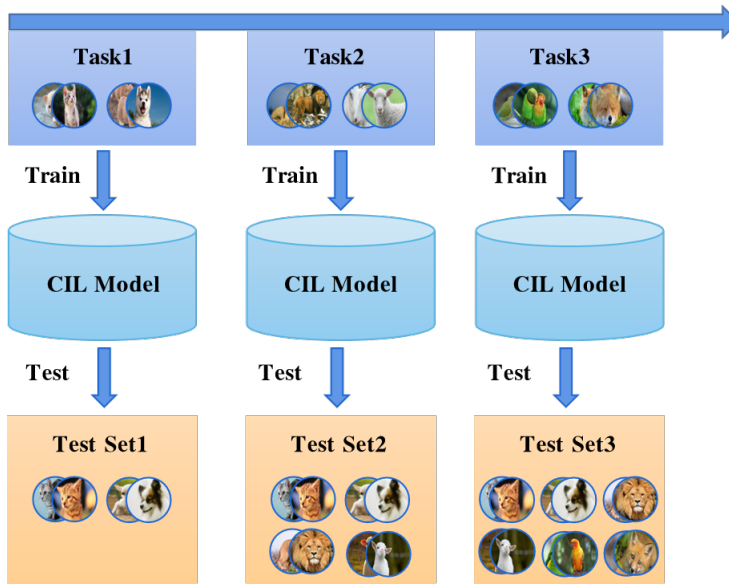

Figure 1. Based on the setup of the class incremental learning environment, data is gradually introduced by category, while the model needs to gradually learn and categorize all the categories.

egy, we can not only improve the model's ability to process complex images in the real world but also enhance its adaptability in various scenarios. Most importantly, the method helps to solve the catastrophic forgetting problem, allowing the model to better retain useful knowledge previously learned when learning new tasks. Meanwhile, in order to better transfer the knowledge of selected old category samples to the model when a new task arrives, and thus further cope with the catastrophic forgetting problem, we introduce the technique of attentional distillation. Specifically, through attentional distillation, we incorporate the predictions from the teacher's model as additional guidance during student model training. This helps to retain important information from past tasks, fighting the forgetting phenomenon.

In addition, in online data streams, the semantic information in the data streams is often not fully utilized due to the constant changes and rapid emergence of information thus leading to reduced model performance. To address this issue, we proposed the ADVC. The core idea of this strategy is to transform the incoming images into different view pairs, extract the key features by using the attention mechanism [18, 19], and then prompt the model to better understand the semantic content of the images by maximizing the mutual information, so as to improve the performance of the model in the online data stream.

In summary, our contributions are as follows:

- To ensure the effectiveness of the replay strategy, we proposed a two-stage sampling method. The goal of this method is to select important and high-quality samples for replay, thereby significantly improving the model's performance while fighting the catastrophic forgetting problem.
- We proposed an ADVC, aiming to explore crucial semantic information in single-pass data streams and effectively improve the model's classification performance and generalization ability.
- By introducing attention distillation, the student model can be guided by the teacher model to learn similar attention patterns, focusing on important samples and features for the current task, and reducing the forgetting issue concerning past tasks.

A preliminary version of this paper is available at [20]. The new contributions come from two main sources. First, in order to obtain clear and high-quality samples during replay, we proposed a two-stage sampling method. By calculating the average gradient of the image, samples with larger average gradients are selected, where larger average gradients represent higher clarity of the image. Then in the second stage, the sample that causes the largest gradient change to the new sample is further selected for replay. The two-stage sampling method improves the effectiveness of the replay strategy. Second, we introduced an attention distillation technique to align the attention of the student model with that of the teacher model thereby allowing the student model to learn more of the teacher model's knowledge to further mitigate the catastrophic forgetting problem.

2 RELATED WORK

In this section, we first review three typical methods for class-incremental learning and then present research work related to knowledge distillation.

2.1 Class-Incremental Learning

In the realm of incremental learning, two primary task types are Task- Incremental Learning (TIL) [21, 22] and Class-Incremental Learning (CIL) [23, 24, 25, 26, 27]. In TIL, each task has a separate classification head for learning different tasks. In contrast, CIL tasks require the model to maintain and update a unified classification head, making it more challenging. While CIL allows the model to learn from continuously changing data streams, there is an unavoidable issue in this process, namely catastrophic forgetting. To tackle the issue of catastrophic forgetting problem in CIL, researchers have continuously proposed new methods, which can be summarized into three main categories: one is the regularization-based methods [28, 29, 30, 31], which involves adding additional rules to prevent the forgetting of

old knowledge, similar to deliberately revisiting and reinforcing previously learned content when acquiring new knowledge; another is the parameter isolation-based methods [32, 33, 34, 35, 36], where new and old knowledge are kept in separate compartments, akin to storing new items in a new drawer rather than overwriting the old ones; and the last one is the replay-based methods [37, 38, 39, 40, 41, 42], which continually replays old data, similar to repeatedly revisiting past knowledge to prevent its loss. These methods help retain old knowledge while learning new things.

Regularization-based methods consist of two methods: One method involves constraining the impact of parameter updates related to prior tasks by evaluating the correlation between parameters and the prior tasks. For example, Kirkpatrick et al. [7] proposed the EWC (Elastic Weight Consolidation) algorithm, which controls the weight optimization direction by adding regularization to the weights. IMM [43] found the maximum of the Gaussian posterior mixing with the estimated Fisher information matrix. Another method utilizes knowledge distillation techniques to regularize data and preserve information related to old classes. The most representative method of this kind is the Learning without Forgetting (LWF) approach proposed by Li and Hoiem [44]. It introduces distillation loss from the output of the new model into the loss function and then fine-tunes the model on the new task. Parameter isolation-based methods aim to mitigate catastrophic forgetting by differentially isolating the parameters of the new and old models. This approach can be divided into two types: one is the Fixed Architecture (FA) [45, 46] which allows the activation of relevant parameters for each task but does not alter the model's overall structure. It is similar to the construction of a house where the basic structure remains the same, and only the interior decoration is adapted to different needs. The other type is the Dynamic Architecture (DA) [47, 48], which introduces new parameters when adding new tasks while keeping the parameters for old tasks unchanged. On the other hand, replay-based methods allow storing a portion of old data in a buffer for use during the learning of new tasks. However, the effectiveness of this method heavily relies on the choice of sampling strategies. For instance, Chaudhry et al. [49] proposed sampling samples with high predictive entropy and near the decision boundaries, selecting high uncertainty samples as exemplars. Aljundi et al. [50] introduced a greedy sampling strategy for online incremental learning, deciding whether to replace samples by comparing the scores of new samples with candidate replacements. Isele and Cosgun [51] explored the reservoir sampling process to ensure an equal probability of sample retention in memory. Although the above methods have achieved some success, they lack consideration of the quality of the stored samples when selecting samples. For example, although some methods can ensure that the selected samples are representative, these samples may not contain enough detailed information (or the samples may contain other disturbing information), which in turn has an impact on the performance of the model. In our approach, we particularly emphasize the importance of sample quality. In the first stage, we focus on selecting those samples with high clarity through a carefully designed replay mechanism. This selection goes beyond

simply pursuing image clarity; Rather, it ensures the quality of the selected samples while preserving as much details and as much information as possible. Such high-quality samples not only help mitigate the effects of catastrophic forgetting, but also provide richer and more representative training data for the continuous learning of the model. In the second stage, we adopt a more refined strategy aimed at further improving the quality of the selected samples. In this phase, we focus our attention on those samples that can produce the largest gradient changes to the new samples. Such a selection approach not only ensures the clarity of the selected samples, but more importantly captures the key information that has the greatest impact on model learning. With this sophisticated sampling approach, we are able to improve the effectiveness of the replay strategy more efficiently and thus better cope with the catastrophic forgetting problem in class incremental learning.

2.2 Knowledge Distillation

Knowledge Distillation (KD) [52, 53, 54, 55, 56] is a model compression technique used to transfer knowledge from a larger or well-trained network (teacher) to a more compact smaller network (student), allowing the student model to learn knowledge similar to the teacher. In our example, the student model represents the current version that is handling the ongoing task, while the teacher model essentially embodies the student's version based on the achievements and knowledge gained from previous tasks. This means that the student model can utilize and depend on previous knowledge when learning something new. The use of feature maps and attention mechanisms in Knowledge Distillation (KD) has been shown to effectively help the student model acquire higher-quality intermediate representations, thereby improving its performance [57]. For instance, Rebuffi et al. [58] proposed iCaRL, which combines knowledge distillation and representation learning, addressing the issue of imbalanced samples between new and old classes by training the feature extractor and classifier separately. Hou et al. [59] introduced three loss functions to mitigate biases caused by imbalanced new and old samples. Wu et al. [60] proposed BiC, which recalibrates output probabilities before applying distillation loss. Douillard et al. [61] presented PODNet, a spatial distillation loss-based approach to counter catastrophic forgetting. Li et al. [62] introduced neural attention distillation as a method for erasing backdoors. Inspired by knowledge distillation, we introduce attention distillation techniques to online class-incremental learning. By distilling the attention maps from the teacher model to the student model, the student gains more crucial knowledge, thereby helping to alleviate catastrophic forgetting.

Unlike traditional class-incremental learning methods, our method carries out multiple considerations. First, critical high-quality samples are selected for replay through a twice-sampling process to avoid retaining interfering information leading to model forgetting. Secondly, to further fight catastrophic forgetting, we introduced attention distillation to enhance the ability of student models to have old knowl-

edge. Finally, we proposed an ADVC aiming to thoroughly explore the semantic information.

3 PROPOSED METHOD

In this section, we start by describing the problem definition. Subsequently, we provide a detailed introduction to our methods, which include two-stage sampling, ADVC, and attention distillation.

3.1 Problem Definition

Based on recent literature [63, 64], we consider a more realistic setting for online class-incremental learning, where the model continuously learns new classes from a non-stationary data stream while retaining knowledge of old classes. Given that the samples within the data stream are encountered only once, it becomes crucial to thoroughly harness the semantic information embedded within it. Formally, we denote the data stream as $D = \{D_1, D_2, \ldots, D_N\}$ over $X \times Y$, where X and Y represent the samples and their labels, respectively, and N is the number of tasks. It is important to note that the classes across different tasks do not overlap. Our objective is to train a new model F_t that can accurately classify all the learned classes. F_t represents the deep image classification model when learning task t. The output of F_t is defined as follows:

$$F_t(x) = \left[F_t^1(x), \ldots, F_t^{n_{t-1}}(x), \ldots, F_t^{n_t}(x)\right]. \tag{1}$$

3.2 Two-Stage Sampling

In class-incremental learning, replay-based methods have achieved state-of-the-art performance. These methods store samples from past tasks in a replay buffer. When new data becomes available, the model performs joint training by sampling from the replay buffer using a specific sampling strategy. Therefore, the key to the success of replay methods lies in the design of the sampling strategy. Figure 2 illustrates the flow of our method. In this paper, we consider a real-world issue where the data stream contains low-quality samples. These samples have lower clarity and may contain noise. Once these samples are selected during the sampling process it will greatly affect the performance of the model. To tackle this problem, we proposed a two-stage sampling strategy.

Assuming the memory size is M. First, we apply the CutMix [65] data augmentation operation to the data in memory. Then in the first stage, we compute the average gradients of samples in the memory and rank them in descending order based on their average gradients. The average gradients are used to represent the image clarity, reflecting the expression capacity of the image details in contrast.

$$\mathbf{g} = |\mathbf{g}_a| + |\mathbf{g}_b|, \tag{2}$$

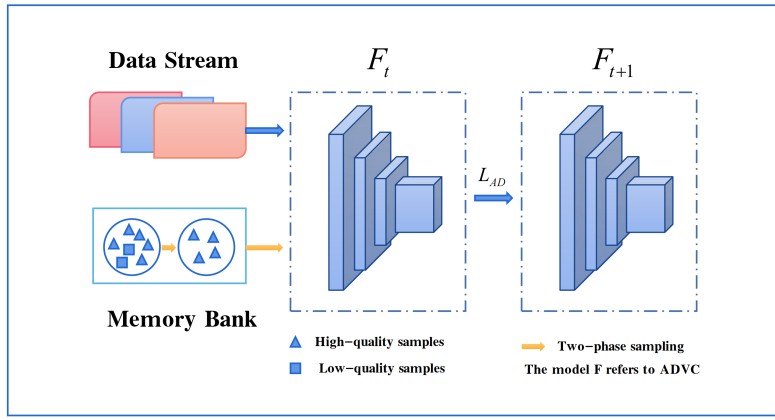

Figure 2. This is the workflow of our method. We will select high-quality samples through a two-phase sampling method in the Memory Bank, and combine them with samples from the data stream to feed into the ADVC network for exploring important semantic information. Then, we use attention distillation to transfer the attention maps from the F_t (current task) to the model for the upcoming new task, enabling it to acquire knowledge from the old tasks.

where g_a denotes the gradient in the horizontal direction and g_b denotes the gradient in the vertical direction. After that, we perform a descending order based on the average gradient:

$$\mathbf{I}' = \text{argsort}(-\mathbf{g}), \tag{3}$$

where *argsort* is a function that sorts the input vector. I denotes the sorted index.

In the second stage, we select the first S samples from memory based on the indexes obtained in the first stage, and for this S samples, we further select the first K samples that cause the maximum gradient change to the new samples in each update of the model, where $K < S < M$. During the training process, x_t represents the new sample, and x_r represents the samples in the memory. We use the new sample x_t to update the model with a learning rate denoted by α.

$$\theta_v = \theta - \alpha \cdot G_\theta(x_t), \tag{4}$$

where $G_\theta(x_t)$ represents the gradient of a new sample x_t with respect to the model parameters θ. Next, we compute the gradient changes using the samples in the memory bank. Firstly, we calculate the gradients of each sample x_r with respect to the virtual parameters θ_v, denoted as $G_{\theta_v}(x_r)$. Then, we calculate the gradients of each sample x_r with respect to the current parameters θ, denoted as $G_\theta(x_r)$. The

difference between the two represents the gradient change:

$$\Delta G(x_r) = G_{\theta_v}(x_r) - G_{\theta}(x_r), \tag{5}$$

$$Score = Sort(\Delta G(x_r); \text{desc}). \tag{6}$$

Finally, we choose the top K samples from all the examples as the samples we want to retrieve. These samples have the largest gradient changes to the model parameters after the virtual update, and we believe that such samples are useful for updating the neural network based on back gradient propagation. Therefore, we can use these samples to guide the training of the model to better utilize the experience in the memory bank.

3.3 Attention-Based Dual-View Consistency

In the setting of online class-incremental learning, image data in a data stream can usually be observed by a model only once, which means that there is a large amount of image information that is underutilized. To overcome this challenge, we proposed the ADVC, which aims to fully exploit the information in online data streams and significantly improve model performance.

An efficient attention mechanism, CBAM [19], is used in our method to generate weights for each input image as a way to guide the model on which regions it should focus.

$$\begin{aligned} \mathbf{F}'_{x^1} &= \mathrm{M}_{\mathbf{c}}(\mathbf{F}_{x^1}) \otimes \mathbf{F}_{x^1}, \\ \mathbf{F}''_{x^1} &= \mathrm{M}_{\mathbf{s}}(\mathbf{F}'_{x^1}) \otimes \mathbf{F}'_{x^1}. \end{aligned} \tag{7}$$

where $\mathbf{F}_{x^1} \in \mathbb{R}^{C\times H\times W}$ is an intermediate feature map of one of the views of F_x. The channel attention map $\mathbf{M_c} \in \mathbb{R}^{C\times 1\times 1}$ is used to enhance channel features. $\mathbf{M_s} \in \mathbb{R}^{1\times H\times W}$ represents the spatial attention map, which accurately identifies important regions. F''_{x^1} represents the final output. The intermediate feature map of the other view of F_x is represented by $\mathbf{F}_{x^2} \in \mathbb{R}^{C\times H\times W}$. Similarly, the output F''_{x^2} can be determined using Equations (7).

As shown in Figure 3, the attention mechanism plays a crucial role in guiding the model to maximize mutual information for key regions from two different perspectives. By maximizing mutual information, our method effectively facilitates information sharing between different views, which helps the model to better understand the correlation between two views in order to utilize the information in the data flow and adapt to the changing data. By making the co-occurrence of events similar to their individual occurrences, the mutual information $I(X_1; X_2)$ can be approximated as follows:

$$I(X_1; X_2) \approx \frac{1}{3}(H(X_1) + H(X_2) + H(X_1, X_2)). \tag{8}$$

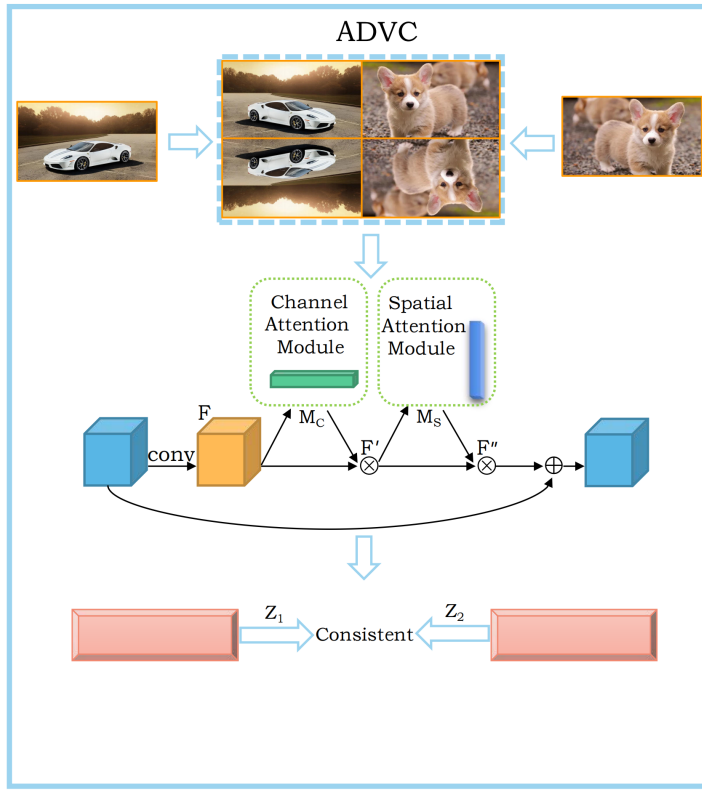

Figure 3. This is the ADVC flow. The left image is a new image incoming from the data stream, and the right image is an old image retrieved from Memory Bank. The two images are transformed into different perspectives, respectively, and the attention mechanism is utilized to fully explore the semantic information contained in the images, eventually improving the consistency of the representations of the image pairs.

For each input image x, we now use the feature F_{x^1} after two rounds of attention transformation, the joint probability matrix $\boldsymbol{P} \in \mathbb{R}^{C \times C}$ can be calculated as:

$$\boldsymbol{P} = \frac{1}{n} \sum_{i=1}^{n} F''\left(\boldsymbol{x}_i^1\right) \cdot F''\left(\boldsymbol{x}_i^2\right)^{\top}. \tag{9}$$

At the time t, there is a batch of n images, with each image having undergone two different transformations, represented as x_i^1 and x_i^2.

Our aim is to maximize $I(z_1; z_2)$, thus the MI loss can be formulated as follows:

$$\mathcal{L}_{MI} = -I\left(\boldsymbol{z}_1; \boldsymbol{z}_2\right), \tag{10}$$

where z_1 and z_2 are used to represent the feature extraction methods for dual-view image pairs. In addition, to ensure the consistency of data representations under different conditions, it is necessary to constrain the difference between the joint distribution and marginal distribution, we use the following loss terms.

$$\mathcal{L}_{DL} = L_1\left(p\left(z_1, z_2\right), p\left(z_1\right)\right) + L_1\left(p\left(z_1, z_2\right), p\left(z_2\right)\right), \tag{11}$$

where L_1 denotes Mean Absolute Error (MAE) loss. Through the attention mechanism and maximizing the mutual information between view pairs, The model will effectively explore the critical semantic information within the data stream.

3.4 Attention Distillation

In class-incremental learning, distillation is a commonly used technique that works like a teacher teaching a student. Specifically, we have an old model (the teacher) that has already learned a lot. Now, we want to train a new model (the student) to learn some new things, but we are concerned that the new model might forget what the old model already knows. So, the distillation technique involves having the teacher provide the student with some guidance about the old knowledge, rather than just giving the student the correct answers. This way, the student can learn new things while retaining and inheriting the knowledge from the teacher, thus preventing catastrophic forgetting.

To achieve this goal, we introduce attention distillation[62] in the context of replay-based class-incremental learning to further alleviate forgetting issues. Specifically, in attention distillation, we train a student network by utilizing the spatial attention maps of the teacher network (computed using attention mapping functions). This allows the student network to extract important information (i.e., neurons that are crucial for old classes) from the already trained teacher model on old classes, enabling it to focus on both new and old class information simultaneously.

Attention Representation. In our model F, $F^d \in \mathbb{R}^{C\times H\times W}$ represents the activation output of the d^{th} layer, where C, H, and W are the dimensions of channels, height, and width of the activation maps, respectively. Subsequently, we define an attention mapping function named $\mathcal{A}$, which transforms an activation map into an attention representation. Specifically, $\mathcal{A}$ takes a three-dimensional activation map F as input and then flattens it along the channel dimension, resulting in a two-dimensional tensor as output.

$$\mathcal{A} : R^{C\times H\times W} \rightarrow R^{H\times W}. \tag{12}$$

According to [62], we use an efficient attention operation to realize the transformation of the activation mapping:

$$\mathcal{A}^p_{\text{sum}}(F^d) = \sum_{i=1}^{C} \left|F_i^d\right|^p, \tag{13}$$

where F_i^d is the activation map of the i^{th} channel, and $p > 1$. The spatial mapping method $\mathcal{A}\text{sum}^p(F^d)$ places greater emphasis on areas where neurons have higher activity levels, which are considered essential for the network. Moreover, as the parameter p increases, the model increasingly prioritizes regions with the highest activity, deeming them more crucial for the task.

Attention Distillation Loss. In our method, we utilize a predefined attention operation formula to calculate the attention representation of the network. The technique of attention representation reveals the network's focus on different parts when processing input data. The teacher network is a network that has already learned a lot, and its knowledge remains constant throughout the entire learning process. The distillation loss at the d^{th} layer of the network takes into account the areas that both the teacher and the student pay attention to when processing data, aiming to assist the student in learning more effectively.

$$\mathcal{L}_{\text{AD}}\left(F_T^d, F_S^d\right) = \left\| \frac{\mathcal{A}(F_T^d)}{\left\|\mathcal{A}(F_T^d)\right\|_2} - \frac{\mathcal{A}(F_S^d)}{\left\|\mathcal{A}(F_S^d)\right\|_2} \right\|_2, \tag{14}$$

where F_T^d represents the activation of the teacher model (old model) at the d^{th} layer, and F_S^d represents the activation of the student model (new model) at the d^{th} layer. $\mathcal{A}()$ is the function used to compute the attention maps, and $\|\cdot\|_2$ denotes the L_2 norm. The attention distillation loss drives the student network to learn similar attention patterns as the teacher network. The attention information from the teacher model reflects its understanding of the old classes. By learning this attention information, the student model can acquire memories of the old classes, thereby mitigating catastrophic forgetting issues. The overall loss function L can be expressed as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{MI} + \lambda_3 \mathcal{L}_{DL} + \beta \cdot \mathcal{L}_{AD}, \tag{15}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss, and λ_1, λ_2, and λ_3 are the balancing coefficients for the four types of losses. β is a hyperparameter controlling the strength of the attention distillation.

4 EXPERIMENT RESULTS AND DISCUSSIONS

In this section, we first review the benchmark datasets used in the study. Three representative datasets were carefully selected for meaningful comparisons. Subsequently, we present the evaluation metrics used. In our experiments, we use average accuracy and average forgetting rate as evaluation metrics to objectively assess the performance of the various methods. Subsequently, we present the baseline methods adopted for comparison with our method. Finally, we not only comprehensively showcase the experimental and ablation study results but also conduct in-depth analyses of these results.

4.1 Datasets

In our research, we selected three widely used datasets in the field of class incremental learning to ensure the generality and comparability of our experimental results.

- The Split CIFAR-10 dataset provides us with a relatively simple scenario, consisting of 10 classes with 5 000 training samples and 1 000 test samples per class. By splitting the CIFAR-10 [66] dataset into 5 small datasets, each involving 2 classes, we can simulate relatively straightforward class-incremental learning situations commonly encountered in real-world applications.
- The CIFAR-100 dataset presents higher challenges, as it includes 100 classes, with each class containing 600 images. By dividing the CIFAR-100 [66] dataset into 10 tasks, each containing 10 disjoint sub-datasets representing individual classes, we effectively simulate class-incremental learning processes on complex tasks.
- The Mini-ImageNet dataset serves as a large-scale dataset, encompassing 100 classes with a total of 60 000 color images, and 600 samples per class. By dividing the Mini-ImageNet [67] dataset into 10 subsets, each containing 10 non-overlapping classes, we face the more challenging task of large-scale incremental class learning.

4.2 Metrics

When evaluating the performance of online class-incremental learning models, common metrics mainly include the average accuracy rate and the rate of forgetting old tasks.

Average accuracy is a metric used to reflect the overall performance of the incremental learning model. One common method for estimating average accuracy was proposed by [68], denoted as A_t. Specifically, when the model completes training on task i and performs classification on the test set of task j, it is represented as $a_{ij} \in [0, 1]$. The formula for calculating the average accuracy A_t for task t is as follows:

$$\text{Average Accuracy}\,(A_t) = \frac{1}{t}\sum_{j=1}^{t} a_{t,j}. \tag{16}$$

Another important evaluation metric is the average forgetting rate, which is used to measure how much the model forgets about old tasks. [46] introduced the forgetting rate F_t to measure how well the model forgets previous tasks on task t. The formula for calculating the average forgetting rate is as follows:

$$\text{Average Forgetting}\,(F_t) = \frac{1}{t-1}\sum_{j=1}^{t-1} f_{t,j}, \tag{17}$$

where $f_{i,j} = \max_{l \in \{1,\ldots,i-1\}} a_{l,j} - a_{i,j}$.

4.3 Baselines

In our experiments, we selected several popular class-incremental learning methods as our baselines to comprehensively compare with the proposed method. These baseline methods include:

- ER (Experience Replay) [68]: This method adopts the reservoir sampling strategy in Memory Retrieval to randomly sample and update the memory. During the learning process, it mitigates the forgetting problem in class-incremental learning by saving historical experiences and training the model with randomly sampled samples from the memory.
- MIR (Maximally Interfered Retrieval) [69]: The MIR method selects old samples that contribute the most to the increase in loss according to the estimated parameters of the current task. The purpose of this method is to prevent the model from excessively focusing on new tasks and forgetting knowledge of old tasks.
- GDumb (Greedy sampler and Dumb learner) [70]: GDumb is a simple yet efficient online incremental learning model. Its key feature is to update the cache in a greedy manner and train the model from scratch using the data inside the cache.
- DER++ (Dark Experience Replay) [71]: DER++ utilizes knowledge distillation to retain past experiences and alleviate the forgetting problem.
- GSS (Gradient-based Sample Selection) [50]: GSS aims to increase the gradient diversity of samples in the memory.
- ASER (Adversarial Shapley Value Experience Replay) [72]: ASER uses Shapley Value for Memory Retrieval and Memory Update.
- DVC (Dual View Consistency) [73]: The DVC method retrieves MGI using Maximum Gradient Interference for Memory Retrieval and maximizes mutual information to explore semantic information in single-pass data streams, thereby enhancing the model's learning capability for new tasks.
- AOCIL (Online Class-Incremental Learning Based on Attention) [20]: AOCIL explores important semantic information in the data stream by augmenting the data in the memory bank and through the attention mechanism.

We consider the above class-incremental learning methods as our baselines, and by comparing with them, we validate the effectiveness and advantages of our proposed method in various scenarios. Through sufficient experimental evaluations, we contribute new insights to the field of class-incremental learning research.

4.4 Implementation Details

According to the existing methods [71, 72], we use ResNet18 as the backbone model for all datasets and train the network using stochastic gradient descent with a learning rate of 0.1. The model receives batch sizes of 10 from the data stream for each training iteration, and the batch size K for memory retrieval is also set to 10. Additionally, to perform memory retrieval, we set the number of candidate samples S to 50. The parameter p in Equation (13) is set to 2.

As for the hyperparameter β in Equation (15), we set it to 10^3 divided by the number of elements in the attention map and the batch size per layer. For CIFAR-10, $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 2$. For CIFAR-100 and Mini-ImageNet, $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 4$.

4.5 Results

We compare our proposed method with the baseline method on three data, cifar10, cifar100 and mini-imageNet. From the experimental results, our method achieves advanced results in terms of average accuracy in Table 1 and average forgetting rate in Table 2.

Method	Mini-ImageNet			CIFAR-100			CIFAR-10		
	M = 1 K	M = 2 K	M = 5 K	M = 1 K	M = 2 K	M = 5 K	M = 0.2 K	M = 0.5 K	M = 1 K
ER	10.2 ± 0.5	12.9 ± 0.8	16.4 ± 0.9	11.6 ± 0.5	15.0 ± 0.5	20.5 ± 0.8	23.2 ± 1.0	31.2 ± 1.4	39.7 ± 1.3
MIR	10.1 ± 0.6	14.2 ± 0.9	18.5 ± 1.0	11.3 ± 0.3	15.1 ± 0.3	22.2 ± 0.7	24.6 ± 0.6	32.5 ± 1.5	42.8 ± 1.4
GSS	9.3 ± 0.8	14.1 ± 1.1	15.0 ± 1.1	9.7 ± 0.2	12.4 ± 0.6	16.8 ± 0.8	23.0 ± 0.9	28.3 ± 1.7	37.1 ± 1.6
GDumb	7.3 ± 0.3	11.4 ± 0.2	19.5 ± 0.5	10.0 ± 0.2	13.3 ± 0.4	19.2 ± 0.4	26.6 ± 1.0	31.9 ± 0.9	37.5 ± 1.1
DER++	10.9 ± 0.6	15.0 ± 0.7	17.4 ± 1.5	11.8 ± 0.4	15.7 ± 0.5	20.8 ± 0.8	28.1 ± 1.2	35.4 ± 1.3	42.8 ± 1.9
ASER	11.5 ± 0.6	13.5 ± 0.8	17.8 ± 1.0	14.3 ± 0.5	17.8 ± 0.5	22.8 ± 1.0	29.6 ± 1.0	38.2 ± 1.0	45.1 ± 2.0
DVC	15.4 ± 0.7	17.2 ± 0.8	19.1 ± 0.9	19.7 ± 0.7	22.1 ± 0.9	24.1 ± 0.8	45.4 ± 1.4	50.6 ± 2.9	52.1 ± 2.5
AOCIL	17.3 ± 0.8	19.4 ± 1.2	23.6 ± 0.9	19.4 ± 1.2	21.8 ± 0.8	24.5 ± 0.3	45.1 ± 1.0	50.8 ± 0.7	54.7 ± 0.5
ADOCIL	$\mathbf{20.7 \pm 0.6}$	$\mathbf{22.4 \pm 1.5}$	$\mathbf{25.2 \pm 0.5}$	$\mathbf{23.3 \pm 1.7}$	$\mathbf{25.8 \pm 0.4}$	$\mathbf{28.2 \pm 0.8}$	$\mathbf{48.8 \pm 1.7}$	$\mathbf{54.6 \pm 0.9}$	$\mathbf{57.1 \pm 0.8}$

Table 1. Average Accuracy (higher values indicate better performance). M represents the memory buffer size. Results in bold represent the best outcomes. All figures represent the average of 15 iterations.

Method	Mini-ImageNet			CIFAR-100			CIFAR-10		
	M = 1 K	M = 2 K	M = 5 K	M = 1 K	M = 2 K	M = 5 K	M = 0.2 K	M = 0.5 K	M = 1 K
ER	32.7 ± 0.9	29.1 ± 0.7	26.0 ± 1.0	39.1 ± 0.9	34.6 ± 0.9	30.6 ± 0.9	60.9 ± 1.0	50.2 ± 2.5	39.5 ± 1.6
MIR	31.5 ± 1.2	25.6 ± 1.1	20.4 ± 1.0	39.5 ± 0.6	33.3 ± 0.8	28.3 ± 0.7	61.8 ± 1.0	51.5 ± 1.4	38.0 ± 1.5
GSS	33.5 ± 0.8	28.0 ± 0.7	27.5 ± 1.2	38.2 ± 0.7	34.3 ± 0.6	30.2 ± 0.8	62.2 ± 1.3	55.3 ± 1.3	44.9 ± 1.4
DER++	33.8 ± 0.8	28.6 ± 0.8	27.1 ± 1.3	41.9 ± 0.6	36.7 ± 0.5	33.5 ± 0.8	55.9 ± 1.8	45.0 ± 1.0	34.6 ± 2.8
ASER	33.8 ± 1.3	30.5 ± 1.3	25.1 ± 0.8	43.0 ± 0.5	37.9 ± 0.6	29.6 ± 0.9	56.4 ± 1.6	47.5 ± 1.3	39.6 ± 2.0
DVC	25.1 ± 0.7	23.1 ± 0.7	21.9 ± 0.8	30.6 ± 0.7	27.8 ± 1.0	26.1 ± 0.5	27.2 ± 2.5	$\mathbf{21.3 \pm 3.1}$	19.7 ± 2.9
AOCIL	24.6 ± 0.9	22.6 ± 0.6	17.8 ± 1.2	29.2 ± 0.5	26.7 ± 0.8	28.7 ± 1.0	$\mathbf{25.1 \pm 1.0}$	30.5 ± 1.7	24.6 ± 0.5
ADOCIL	$\mathbf{19.3 \pm 1.2}$	$\mathbf{17.5 \pm 0.9}$	$\mathbf{14.8 \pm 1.9}$	$\mathbf{22.1 \pm 0.7}$	$\mathbf{19.4 \pm 0.6}$	$\mathbf{16.8 \pm 1.3}$	30.2 ± 1.5	23.1 ± 1.2	$\mathbf{17.4 \pm 0.6}$

Table 2. Average Forgetting (lower values indicate better performance). M represents the memory buffer size. Results in bold represent the best outcomes. All figures represent the average of 15 iterations.

Results of average accuracy. From the results of average accuracy, our method has shown significant improvements on all three datasets, regardless of the memory size. For the CIFAR-10 dataset, our method achieved an average improvement of 4.1 %. For the CIFAR-100 dataset, the average improvement was 3.8 %. The most significant improvement was observed on the Mini-ImageNet dataset, with an average accuracy increase of 5.5 %, especially at M = 1 K, where the improvement reached 6.1 %.

We believe that such a significant improvement is due to the effectiveness of our method. Two-stage sampling and attention distillation enhance model performance by fighting catastrophic forgetting. And ADVC fully explores useful semantic information in the data stream during the training process of the model, which improves the model's adaptability and generalization to new tasks.

Results of average forgetting rate. Regarding the average forgetting rate, our method also exhibits significant improvements. For the CIFAR-10 dataset, our method did not achieve superior results at M = 0.2 K and M = 0.5 K. The reason for this outcome is that the introduced attention mechanism tends to overly focus on a small number of samples when the memory size is small, leading to an increase in forgetting rate for other samples. However, at M = 1 K, our method outperformed the baseline by reducing the forgetting rate by 2.3 %. On the CIFAR-100 and Mini-ImageNet datasets, our method reduced the average forgetting rate by 8.7 % and 6.1 %, respectively. This clearly indicates that our method better preserves previous information, thereby alleviating the catastrophic forgetting problem.

Figure 4 and Figure 5 show that our method consistently outperforms other baseline methods. As the number of tasks increases, our sample selection method is able to filter out more important and high-quality samples during replay and further fights the catastrophic forgetting problem by distilling the attention map to the student model so that the student model retains more critical information about the old classes. Meanwhile, our ADVC is able to explore important semantic information in the data stream, enabling our method to achieve higher accuracy. The outstanding performance across various tasks validates our method's ability to preserve previously acquired knowledge while efficiently adapting to new tasks, positioning our method as a powerful solution for mitigating catastrophic forgetting in continuous learning environments.

4.6 Ablation Study

Our method mainly consists of three components: Two-Stage Sampling, ADVC, and Attention Distillation. Among them, the role of ADVC is to better explore information in the data stream to improve the model's performance, while Two-Stage Sampling and Attention Distillation mainly aim to mitigate catastrophic forgetting. Therefore, in Table 2, we conduct an ablation study on the average accuracy

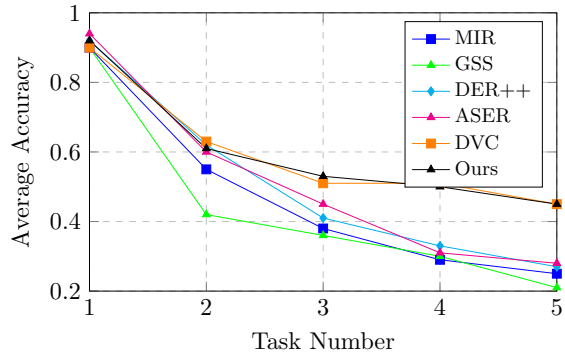

Figure 4. Observe the average accuracy of each task on the CIFAR-10 dataset with M = 200

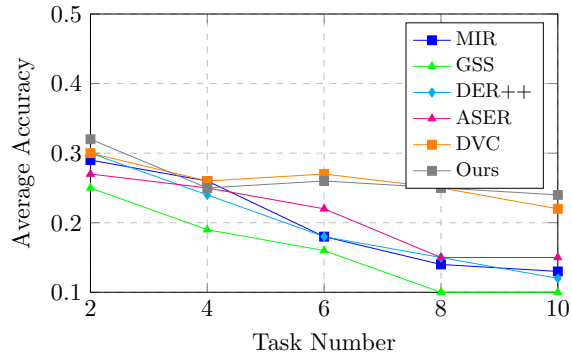

Figure 5. Observe the average accuracy of each task on the CIFAR-100 dataset with M = 1 000

improvement brought by each component, and in Table 4, we demonstrate the improvements in average forgetting rate due to Two-Stage Sampling and Attention Distillation. As shown in Table 3, each component of our method contributes to the improvement of model performance. Specifically, Two-Stage Sampling demonstrates the most significant improvement, especially at M = 2 K, where it outperforms the baseline by 3.3 %. We believe that Two-Stage Sampling helps us select high-quality samples with significant impacts. ADVC and Attention Distillation also exhibit outstanding performance, as both mechanisms enhance the model's representation capacity and generalization performance, improving the model's ability to recognize and utilize crucial samples.

Table 4 presents the results of the ablation study on the average forgetting rate for Two-Stage Sampling and Attention Distillation. From the results in the table, both methods significantly reduce the average forgetting rate. Especially noteworthy is the remarkable improvement achieved by Attention Distillation, with

Method	M = 1 K	M = 2 K	M = 5 K
Baseline	19.7 ± 0.7	22.1 ± 0.9	24.1 ± 0.8
Baseline + ADVC	20.8 ± 0.5	23.1 ± 1.7	26.0 ± 0.6
Baseline + Attention distillation	22.1 ± 1.6	23.9 ± 0.4	26.7 ± 1.4
Baseline + Two-stage sampling	22.7 ± 1.4	25.4 ± 1.6	27.6 ± 1.3
Baseline + ADVC + Attention distillation + Two-stage sampling	$\mathbf{23.3 \pm 1.7}$	$\mathbf{25.8 \pm 0.4}$	$\mathbf{28.2 \pm 0.8}$

Table 3. Average Accuracy (higher values indicate better performance). Ablation studies on CIFAR-100. "Baseline" denotes the model employing the DVC method. All figures represent the average of 15 iterations.

Method	M = 1 K	M = 2 K	M = 5 K
Baseline	30.6 ± 0.7	27.8 ± 1.0	26.1 ± 0.5
Baseline + Two-stage sampling	21.4 ± 1.5	$\mathbf{18.6 \pm 1.0}$	17.7 ± 1.0
Baseline + Attention distillation	$\mathbf{20.7 \pm 0.6}$	19.6 ± 1.5	$\mathbf{17.1 \pm 0.5}$

Table 4. Average Forgetting (lower values indicate better performance). Ablation studies on CIFAR-100. "Baseline" denotes the model employing the DVC method. All figures represent the average of 15 iterations.

an average forgetting rate reduction of 9.0 %. This significant improvement is mainly attributed to the alignment of attention distributions between the student model and the teacher model through attention distillation. In incremental learning tasks, where data and tasks continuously change, the model needs to quickly adapt to new tasks while retaining knowledge of previous tasks. Attention distillation allows the student model to focus more on samples and features relevant to the current task while reducing forgetting of the previous tasks.

4.7 ADOCIL Versus AOCIL

The preliminary version of our proposed method is AOCIL [20] and the improvements are described in Section 1. To demonstrate the enhanced effectiveness of our improved method over the previous one, we compare the outcomes of ADOCIL and AOCIL, as depicted in Tables 1 and 2.

Based on the information provided in the table, we observe the average accuracy of AOCIL and ADOCIL across various datasets (Mini-ImageNet, CIFAR-100, CIFAR-10). Among them, the results of ADOCIL outperform those of AOCIL on all datasets. This is mainly attributed to our proposed two-stage sampling strategy. This strategy mitigates the catastrophic forgetting problem and improves the performance of the model by filtering out high-quality samples and further selecting among these high-quality samples the ones that cause the largest gradient changes to the new samples for replay.

As illustrated in Table 2, the enhancement effect of ADOCIL is more pronounced on the Mini-ImageNet and CIFAR-100 datasets compared to the CIFAR-10 dataset. This is due to the fact that the attentional distillation technique we introduced can alleviate the phenomenon of forgetting the history task. By mapping the attention map of the teacher model to the student model, the attention distribution of the student model can be made closer to that of the teacher model, thus retaining more information about the historical task.

5 CONCLUSION

In this paper, we introduce the ADOCIL framework, comprising three main components. Among these, the two-stage sampling method facilitates the prioritization of crucial samples among high-quality ones. Specifically, we select samples based on their larger average gradients and the extent to which new incoming samples perturb the network's gradients. We believe that such samples can retain sufficient information from old classes, thus alleviating the problem of catastrophic forgetting. Furthermore, to explore crucial information in the data stream, we proposed the ADVC. The method utilizes an attention mechanism to focus on important information and then maximizes the mutual information between them to enhance the model's understanding of the data. Most importantly, we introduced an Attention Distillation technology, aligning the attention distribution of the student model with that of the teacher model. This allows the student model to pay more attention to samples and features relevant to the current task, ultimately fighting catastrophic forgetting. Extensive experiments validate the superior performance of our method in terms of average accuracy and average forgetting rate.

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability: Data will be made available on request.

Acknowledgement: This work was supported by the Natural Science Foundation of Shandong Province, China, under Grant Nos. ZR2020MF041 and ZR2022-MF237, and the National Natural Science Foundation of China under Grant No. 11901325.

CRediT authorship contribution statement:

- Jinyong Cheng: conceptualization, methodology, software, validation, visualization, writing – original draft, writing – review and editing;
- Mengyun Chen: software, validation, visualization, writing – original draft, writing – review and editing;
- Baoyu Du: investigation, supervision;
- Min Guo: conceptualization, writing – review, editing and supervision.

REFERENCES

[1] LeCun, Y.—Bengio, Y.—Hinton, G.: Deep Learning. Nature, Vol. 521, 2015, No. 7553, pp. 436–444, doi: 10.1038/nature14539.

[2] Kamilaris, A.—Prenafeta-Boldú, F. X.: Deep Learning in Agriculture: A Survey. Computers and Electronics in Agriculture, Vol. 147, 2018, pp. 70–90, doi: 10.1016/j.compag.2018.02.016.

[3] Sarker, I. H.: Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Computer Science, Vol. 2, 2021, No. 6, Art. No. 420, doi: 10.1007/s42979-021-00815-1.

[4] Gomes, H. M.—Barddal, J. P.—Enembreck, F.—Bifet, A.: A Survey on Ensemble Learning for Data Stream Classification. ACM Computing Surveys (CSUR), Vol. 50, 2017, No. 2, Art. No. 23, doi: 10.1145/3054925.

[5] Golab, L.—Özsu, M. T.: Issues in Data Stream Management. ACM SIGMOD Record, Vol. 32, 2003, No. 2, pp. 5–14, doi: 10.1145/776985.776986.

[6] French, R. M.: Catastrophic Forgetting in Connectionist Networks. Trends in Cognitive Sciences, Vol. 3, 1999, No. 4, pp. 128–135, doi: 10.1016/s1364-6613(99)01294-2.

[7] Kirkpatrick, J.—Pascanu, R.—Rabinowitz, N.—Veness, J.—Desjardins, G.—Rusu, A. A.—Milan, K.—Quan, J.—Ramalho, T.—Grabska-Barwinska, A.—Hassabis, D.—Clopath, C.—Kumaran, D.—Hadsell, R.: Overcoming Catastrophic Forgetting in Neural Networks. Proceedings of the National Academy of Sciences, Vol. 114, 2017, No. 13, pp. 3521–3526, doi: 10.1073/pnas.1611835114.

[8] Kemker, R.—McClure, M.—Abitino, A.—Hayes, T.—Kanan, C.: Measuring Catastrophic Forgetting in Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, No. 1, doi: 10.1609/aaai.v32i1.11651.

[9] Xie, Z.—He, F.—Fu, S.—Sato, I.—Tao, D.—Sugiyama, M.: Artificial Neural Variability for Deep Learning: On Overfitting, Noise Memorization, and Catastrophic Forgetting. Neural Computation, Vol. 33, 2021, No. 8, pp. 2163–2192, doi: 10.1162/neco_a_01403.

[10] Fu, Z.—Wang, Z.—Xu, X.—Li, D.—Yang, H.: Knowledge Aggregation Networks for Class Incremental Learning. Pattern Recognition, Vol. 137, 2023, Art. No. 109310, doi: 10.1016/j.patcog.2023.109310.

[11] Wang, Z.—Xu, L.—Qiu, Z.—Wu, Q.—Meng, F.—Li, H.: GFR: Generic Feature Representations for Class Incremental Learning. Neurocomputing, Vol. 548, 2023, Art. No. 126410, doi: 10.1016/j.neucom.2023.126410.

[12] Simon, C.—Koniusz, P.—Harandi, M.: On Learning the Geodesic Path for Incremental Learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1591–1600, doi: 10.1109/CVPR46437.2021.00164.

[13] Belouadah, E.—Popescu, A.—Kanellos, I.: A Comprehensive Study of Class Incremental Learning Algorithms for Visual Tasks. Neural Networks, Vol. 135, 2021, pp. 38–54, doi: 10.1016/j.neunet.2020.12.003.

[14] Zhou, D. W.—Wang, Q. W.—Qi, Z. H.—Ye, H. J.—Zhan, D. C.—Liu, Z.: Deep Class-Incremental Learning: A Survey. CoRR, 2023, doi: 10.48550/arXiv.2302.03648.

[15] De Lange, M.—Aljundi, R.—Masana, M.—Parisot, S.—Jia, X.—Leonardis, A.—Slabaugh, G.—Tuytelaars, T.: A Continual Learning Survey: Defying Forgetting in Classification Tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, 2022, No. 7, pp. 3366–3385, doi: 10.1109/TPAMI.2021.3057446.

[16] Fan, L.—Zhang, F.—Fan, H.—Zhang, C.: Brief Review of Image Denoising Techniques. Visual Computing for Industry, Biomedicine, and Art, Vol. 2, 2019, No. 1, Art. No. 7, doi: 10.1186/s42492-019-0016-7.

[17] Dodge, S.—Karam, L.: Understanding How Image Quality Affects Deep Neural Networks. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, 2016, pp. 1–6, doi: 10.1109/QoMEX.2016.7498955.

[18] Hu, J.—Shen, L.—Sun, G.: Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[19] Woo, S.—Park, J.—Lee, J. Y.—Kweon, I. S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11211, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[20] Du, B.—Wei, Z.—Cheng, J.—Lv, G.—Dai, X.: Online Class-Incremental Learning in Image Classification Based on Attention. In: Liu, Q., Wang, H., Ma, Z. et al. (Eds.): Pattern Recognition and Computer Vision (PRCV 2023). Springer, Singapore, Lecture Notes in Computer Science, Vol. 14431, 2024, pp. 487–499, doi: 10.1007/978-981-99-8540-1_39.

[21] Feng, F.—Chan, R. H. M.—Shi, X.—Zhang, Y.—She, Q.: Challenges in Task Incremental Learning for Assistive Robotics. IEEE Access, Vol. 8, 2020, pp. 3434–3441, doi: 10.1109/ACCESS.2019.2955480.

[22] Belouadah, E.—Popescu, A.: IL2M: Class Incremental Learning with Dual Memory. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 583–592, doi: 10.1109/ICCV.2019.00067.

[23] Wu, T. Y.—Swaminathan, G.—Li, Z.—Ravichandran, A.—Vasconcelos, N.—Bhotika, R.—Soatto, S.: Class-Incremental Learning with Strong Pre-Trained Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9591–9600, doi: 10.1109/CVPR52688.2022.00938.

[24] Liu, Y.—Su, Y.—Liu, A. A.—Schiele, B.—Sun, Q.: Mnemonics Training: Multi-Class Incremental Learning Without Forgetting. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12242–12251, doi: 10.1109/CVPR42600.2020.01226.

[25] Zhou, D. W.—Wang, F. Y.—Ye, H. J.—Ma, L.—Pu, S.—Zhan, D. C.: Forward Compatible Few-Shot Class-Incremental Learning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9036–9046,

doi: 10.1109/CVPR52688.2022.00884.

[26] Chen, Y.—Li, Z.—Hu, Z.—Vasconcelos, N.: Taxonomic Class Incremental Learning. CoRR, 2023, doi: 10.48550/arXiv.2304.05547.

[27] Liu, Y.—Li, Y.—Schiele, B.—Sun, Q.: Online Hyperparameter Optimization for Class-Incremental Learning. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, No. 7, pp. 8906–8913, doi: 10.1609/aaai.v37i7.26070.

[28] Aljundi, R.—Babiloni, F.—Elhoseiny, M.—Rohrbach, M.—Tuytelaars, T.: Memory Aware Synapses: Learning What (Not) to Forget. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11207, 2018, pp. 144–161, doi: 10.1007/978-3-030-01219-9_9.

[29] Ritter, H.—Botev, A.—Barber, D.: Online Structured Laplace Approximations for Overcoming Catastrophic Forgetting. Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018, pp. 3738–3748, https://proceedings.neurips.cc/paper_files/paper/2018/file/f31b20466ae89669f9741e047487eb37-Paper.pdf.

[30] Zenke, F.—Poole, B.—Ganguli, S.: Continual Learning Through Synaptic Intelligence. In: Precup, D., Teh, Y. W. (Eds.): Proceedings of the 34th International Conference on Machine Learning. PMLR, Vol. 70, 2017, pp. 3987–3995, https://proceedings.mlr.press/v70/zenke17a.html.

[31] Wang, S.—Li, X.—Sun, J.—Xu, Z.: Training Networks in Null Space of Feature Covariance for Continual Learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 184–193, doi: 10.1109/CVPR46437.2021.00025.

[32] Lee, S.—Ha, J.—Zhang, D.—Kim, G.: A Neural Dirichlet Process Mixture Model for Task-Free Continual Learning. CoRR, 2020, doi: 10.48550/arXiv.2001.00689.

[33] Mallya, A.—Lazebnik, S.: PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7765–7773, doi: 10.1109/CVPR.2018.00810.

[34] Yoon, J.—Yang, E.—Lee, J.—Hwang, S. J.: Lifelong Learning with Dynamically Expandable Networks. CoRR, 2017, doi: 10.48550/arXiv.1708.01547.

[35] Yoon, J.—Kim, S.—Yang, E.—Hwang, S. J.: Scalable and Order-Robust Continual Learning with Additive Parameter Decomposition. CoRR, 2019, doi: 10.48550/arXiv.1902.09432.

[36] Saha, G.—Garg, I.—Ankit, A.—Roy, K.: SPACE: Structured Compression and Sharing of Representational Space for Continual Learning. IEEE Access, Vol. 9, 2021, pp. 150480–150494, doi: 10.1109/ACCESS.2021.3126027.

[37] Chaudhry, A.—Gordo, A.—Dokania, P.—Torr, P.—Lopez-Paz, D.: Using Hindsight to Anchor Past Knowledge in Continual Learning. Proceedings of the AAAI Conference on Artificial Intelligencee, Vol. 35, 2021, No. 8, pp. 6993–7001, doi: 10.1609/aaai.v35i8.16861.

[38] De Lange, M.—Tuytelaars, T.: Continual Prototype Evolution: Learning Online from Non-Stationary Data Streams. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8230–8239, doi: 10.1109/ICCV48922.2021.00814.

[39] Shin, H.—Lee, J. K.—Kim, J.—Kim, J.: Continual Learning with Deep Generative Replay. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., 2017, pp. 2990–2999, https://proceedings.neurips.cc/paper_files/paper/2017/file/0efbe98067c6c73dba1250d2beaa81f9-Paper.pdf.

[40] He, C.—Wang, R.—Shan, S.—Chen, X.: Exemplar-Supported Generative Reproduction for Class Incremental Learning. British Machine Vision Conference 2018 (BMVC 2018), 2018, https://bmva-archive.org.uk/bmvc/2018/contents/papers/0325.pdf.

[41] Kemker, R.—Kanan, C.: FearNet: Brain-Inspired Model for Incremental Learning. CoRR, 2017, doi: 10.48550/arXiv.1711.10563.

[42] Zhu, F.—Zhang, X. Y.—Wang, C.—Yin, F.—Liu, C. L.: Prototype Augmentation and Self-Supervision for Incremental Learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5871–5880, doi: 10.1109/CVPR46437.2021.00581.

[43] Lee, S. W.—Kim, J. H.—Jun, J.—Ha, J. W.—Zhang, B. T.: Overcoming Catastrophic Forgetting by Incremental Moment Matching. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., 2017, pp. 4652–4662, https://proceedings.neurips.cc/paper/2017/file/f708f064faaf32a43e4d3c784e6af9ea-Paper.pdf.

[44] Li, Z.—Hoiem, D.: Learning Without Forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, 2018, No. 12, pp. 2935–2947, doi: 10.1109/TPAMI.2017.2773081.

[45] Fernando, C.—Banarse, D.—Blundell, C.—Zwols, Y.—Ha, D.—Rusu, A. A.—Pritzel, A.—Wierstra, D.: PathNet: Evolution Channels Gradient Descent in Super Neural Networks. CoRR, 2017, doi: 10.48550/arXiv.1701.08734.

[46] Serra, J.—Suris, D.—Miron, M.—Karatzoglou, A.: Overcoming Catastrophic Forgetting with Hard Attention to the Task. In: Dy, J., Krause, A. (Eds.): Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR), Vol. 80, 2018, pp. 4548–4557, https://proceedings.mlr.press/v80/serra18a.html.

[47] Aljundi, R.—Chakravarty, P.—Tuytelaars, T.: Expert Gate: Lifelong Learning with a Network of Experts. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3366–3375, doi: 10.1109/CVPR.2017.753.

[48] Rusu, A. A.—Rabinowitz, N. C.—Desjardins, G.—Soyer, H.—Kirkpatrick, J.—Kavukcuoglu, K.—Pascanu, R.—Hadsell, R.: Progressive Neural Networks. CoRR, 2016, doi: 10.48550/arXiv.1606.04671.

[49] Chaudhry, A.—Dokania, P. K.—Ajanthan, T.—Torr, P. H. S.: Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11215, 2018, pp. 532–547, doi: 10.1007/978-3-030-01252-6_33.

[50] Aljundi, R.—Lin, M.—Goujaud, B.—Bengio, Y.: Gradient Based Sample Selection for Online Continual Learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Curran Associates, Inc., 2019, pp. 11816–11825, https://proceedings.neurips.cc/paper_files/paper/2019/file/e562cd9c0768d5464b64cf61da7fc6bb-Paper.pdf.

[51] Isele, D.—Cosgun, A.: Selective Experience Replay for Lifelong Learning. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, pp. 3302–3309, doi: 10.1609/aaai.v32i1.11595.

[52] Gou, J.—Yu, B.—Maybank, S. J.—Tao, D.: Knowledge A Survey. International Journal of Computer Vision, Vol. 129, 2021, No. 6, pp. 1789–1819, doi: 10.1007/s11263-021-01453-z.

[53] Wang, L.—Yoon, K. J.: Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, 2022, No. 6, pp. 3048–3068, doi: 10.1109/TPAMI.2021.3055564.

[54] Huang, T.—You, S.—Wang, F.—Qian, C.—Xu, C.: Knowledge Distillation from a Stronger Teacher. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.): Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Curran Associates, Inc., 2022, pp. 33716–33727, https://proceedings.neurips.cc/paper_files/paper/2022/file/da669dfd3c36c93905a17ddba01eef06-Paper-Conference.pdf.

[55] Beyer, L.—Zhai, X.—Royer, A.—Markeeva, L.—Anil, R.—Kolesnikov, A.: Knowledge Distillation: A Good Teacher Is Patient and Consistent. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10915–10924, doi: 10.1109/CVPR52688.2022.01065.

[56] Liu, Y.—Cao, J.—Li, B.—Yuan, C.—Hu, W.—Li, Y.—Duan, Y.: Knowledge Distillation via Instance Relationship Graph. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7089–7097, doi: 10.1109/CVPR.2019.00726.

[57] Zagoruyko, S.—Komodakis, N.: Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. CoRR, 2017, doi: 10.48550/arXiv.1612.03928.

[58] Rebuffi, S. A.—Kolesnikov, A.—Sperl, G.—Lampert, C. H.: iCaRL: Incremental Classifier and Representation Learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5533–5542, doi: 10.1109/CVPR.2017.587.

[59] Hou, S.—Pan, X.—Loy, C. C.—Wang, Z.—Lin, D.: Learning a Unified Classifier Incrementally via Rebalancing. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 831–839, doi: 10.1109/CVPR.2019.00092.

[60] Wu, Y.—Chen, Y.—Wang, L.—Ye, Y.—Liu, Z.—Guo, Y.—Fu, Y.: Large Scale Incremental Learning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 374–382, doi: 10.1109/CVPR.2019.00046.

[61] Douillard, A.—Cord, M.—Ollion, C.—Robert, T.—Valle, E.: PODNet: Pooled Outputs Distillation for Small-Tasks Incremental Learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12365, 2020, pp. 86–102, doi: 10.1007/978-3-030-58565-5_6.

[62] Li, Y.—Lyu, X.—Koren, N.—Lyu, L.—Li, B.—Ma, X.: Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. CoRR, 2021, doi: 10.48550/arXiv.2101.05930.

[63] Aljundi, R.—Lin, M.—Goujaud, B.—Bengio, Y.: Online Continual Learning with No Task Boundaries. CoRR, 2019, doi: 10.48550/arXiv.1903.08671.

[64] Mai, Z.—Li, R.—Jeong, J.—Quispe, D.—Kim, H.—Sanner, S.: Online Continual Learning in Image Classification: An Empirical Survey. Neurocomputing, Vol. 469, 2022, pp. 28–51, doi: 10.1016/j.neucom.2021.10.021.

[65] Yun, S.—Han, D.—Chun, S.—Oh, S. J.—Yoo, Y.—Choe, J.: CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6022–6031, doi: 10.1109/ICCV.2019.00612.

[66] Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. 2009, https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

[67] Vinyals, O.—Blundell, C.—Lillicrap, T.—Kavukcuoglu, K.—Wierstra, D.: Matching Networks for One Shot Learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 29 (NIPS 2016). Curran Assocates, Inc., 2016, pp. 3630–3638, https://proceedings.neurips.cc/paper_files/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

[68] Chaudhry, A.—Rohrbach, M.—Elhoseiny, M.—Ajanthan, T.—Dokania, P. K.—Torr, P. H. S.—Ranzato, M.: On Tiny Episodic Memories in Continual Learning. CoRR, 2019, doi: 10.48550/arXiv.1902.10486.

[69] Aljundi, R.—Belilovsky, E.—Tuytelaars, T.—Charlin, L.—Caccia, M.—Lin, M.—Page-Caccia, L.: Online Continual Learning with Maximally Interfered Retrieval. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Curran Associates, Inc., 2019, pp. 11849–11860, https://proceedings.neurips.cc/paper_files/paper/2019/file/15825aee15eb335cc13f9b559f166ee8-Paper.pdf.

[70] Prabhu, A.—Torr, P. H. S.—Dokania, P. K.: GDumb: A Simple Approach That Questions Our Progress in Continual Learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. M. (Eds.): Computer Vision – ECCV 2020. Springer, Cham, Lecture Notes in Computer Science, Vol. 12347, 2020, pp. 524–540, doi: 10.1007/978-3-030-58536-5_31.

[71] Buzzega, P.—Boschini, M.—Porrello, A.—Abati, D.—Calderara, S.: Dark Experience for General Continual Learning: A Strong, Simple Baseline. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., Lin, H. (Eds.): Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Curran Associates, Inc.,

2020, pp. 15920–15930, https://proceedings.neurips.cc/paper_files/paper/2020/file/b704ea2c39778f07c617f6b7ce480e9e-Paper.pdf.

[72] Shim, D.—Mai, Z.—Jeong, J.—Sanner, S.—Kim, H.—Jang, J.: Online Class-Incremental Continual Learning with Adversarial Shapley Value. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, No. 11, pp. 9630–9638, doi: 10.1609/aaai.v35i11.17159.

[73] Gu, Y.—Yang, X.—Wei, K.—Deng, C.: Not Just Selection, But Exploration: Online Class-Incremental Continual Learning via Dual View Consistency. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7432–7441, doi: 10.1109/CVPR52688.2022.00729.

Jinyong Cheng received his M.Sc. degree in computer software and theory from the Qingdao University, Qingdao, China, in 2005. He is currently Associate Professor of the School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences). His research interests include artificial intelligence, image processing, and biomedical information processing. Until now, he has published more than 30 papers in conferences and journals. His research is sponsored by the Natural Science Foundation of Shandong province.

Mengyun Chen earned her B.Sc. degree from the Dezhou University in Dezhou, China in 2023. She is currently pursuing her Master's degree in the School of Computer Science and Technology at the Qilu University of Technology. Her current research interests include computer vision and machine learning.

Baoyu Du received her B.Sc. degree from the ShanDong Jiao-Tong University in Jinan, China in 2021. She is currently working toward her Master's degree at the School of Computer Science and Technology, Qilu University of Technology. Her current research interests include computer vision and machine learning.

Min Guo is currently Senior Experimentalist at the Network Information Center, Qilu University of Technology (Shandong Academy of Sciences). Her research interests include big data, machine learning and artificial intelligence. She holds her Master's degree from the School of Computer Science and Technology of Qufu Normal University, China.