ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888118

Enhancing Data-Free Class-Incremental Learning via Image-Centric Dual Distillation

Feifei Fu, Zhiwu Lu*
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

***Abstract*—In Data-free Class-Incremental Learning (DFCIL), catastrophic forgetting is a significant challenge due to the lack of access to previous task image data. Recent approaches using model inversion have made progress in addressing this issue, yet the suboptimal application of knowledge distillation hampers new task learning, limiting overall model performance. To overcome this, we propose a novel method incorporating image-centric dual distillation, designed to retain more old knowledge while facilitating new knowledge acquisition, thus enhancing DFCIL performance. Specifically, we first introduce a weak-constraint relation distillation strategy to preserve old knowledge while promoting the assimilation of new knowledge by learning the relationships among intra-class samples. Then, to further enhance the preservation of old knowledge and refine the integration of new knowledge, we introduce a low-level feature distillation strategy to retain foundational general knowledge by leveraging semantic information from shallow network layers. Extensive experiments show the effectiveness of our method.**

***Index Terms*—Data-free class-incremental learning, Catastrophic forgetting, Knowledge distillation**

I. Introduction

Class-incremental learning (CIL) [1]–[5] has exhibited substantial advancements in mitigating the catastrophic forgetting [6], [7]. The success mainly relies on the storage of training data from previous tasks. However, this has a large demand for storage, and raises data privacy concerns in many practical applications. In view of the above issues, the data-free CIL (DFCIL) [8]–[14] has garnered significant interest among researchers since there is no need to store the data of previous tasks. Because of this, this setting is more challenging to mitigate the forgetting compared to CIL.

Recent methods for DFCIL using model inversion technique [15] have made great progress in mitigating forgetting. However, they still suffer from shortcomings, particularly the suboptimal application of knowledge distillation (KD) strategies [16]–[18], which impairs the learning of new tasks, thereby limiting overall model performance. For example, the recent work [19] introduces an importance-weighted strategy to preserve important old knowledge. Although this method yields commendable results, it constraints the model performance due to the impairment of learning new tasks caused by strongly regularizing and failing to account for differences between the new and synthesized old data. In a more recent work [20], the authors employ a hard KD strategy on synthesized old data to preserve old knowledge, and employ the relational KD [21] on current new data to mitigate the damage caused by the hard KD. Despite aiding new task learning, this method's drawbacks outweigh its benefits, as hard KD's strict output alignment hinders new task acquisition, limiting overall model performance improvement. Therefore, we need to explore an effective method that can retain more old knowledge while minimizing the damage to the learning of new tasks under the DFCIL setting.

To this end, based on model inversion, we propose a novel method incorporating image-centric dual distillation, i.e., the weak-constraint **R**elation distillation and the low-level **F**eature **D**istillation, called RFD. This method is designed to retain more old knowledge while facilitating new knowledge acquisition thereby enhancing the model performance under the DFCIL setting. Specifically, as shown in Fig. 1, our method introduces the weak-constraint relation distillation (RD) strategy to be applied to the synthesized old data and new data respectively to preserve old knowledge and help new task learning. Further, we implement the low-level feature distillation (FD) strategy to be applied to the new data to enhance the model performance. The complementary application of the weak-constraint RD and low-level FD strategies ensures a comprehensive and robust representation of old and new knowledge within the feature space. Extensive experiments fully demonstrate the effectiveness of our RFD.

The main contributions of our work are: **(1)** We propose a novel method incorporating image-centric dual distillation strategies to retain more old knowledge while facilitating new knowledge acquisition thereby enhancing the model performance under the DFCIL setting. **(2)** We introduce a weak-constraint relation distillation strategy to retain more useful old knowledge while promoting the assimilation of new knowledge by learning the correlated relationships among intra-class image samples. Furthermore, to enhance the preservation of old knowledge and refine the integration of new knowledge, we introduce a low-level feature distillation strategy to dig deep and retain fundamental general knowledge by leveraging semantic information from shallow network layers. **(3)** Extensive experiments on multiple benchmarks demonstrate that our method significantly outperforms the state-of-the-arts under the DFCIL setting.

II. Methodology

A. Problem Definition

In the CIL setting, a set of N tasks, denoted as $T = \{T_1, T_2, ..., T_N\}$, are presented for sequential learning. For each task t ($T_t, t \in [1, N]$), it comprises the task-specific

* Corresponding author.

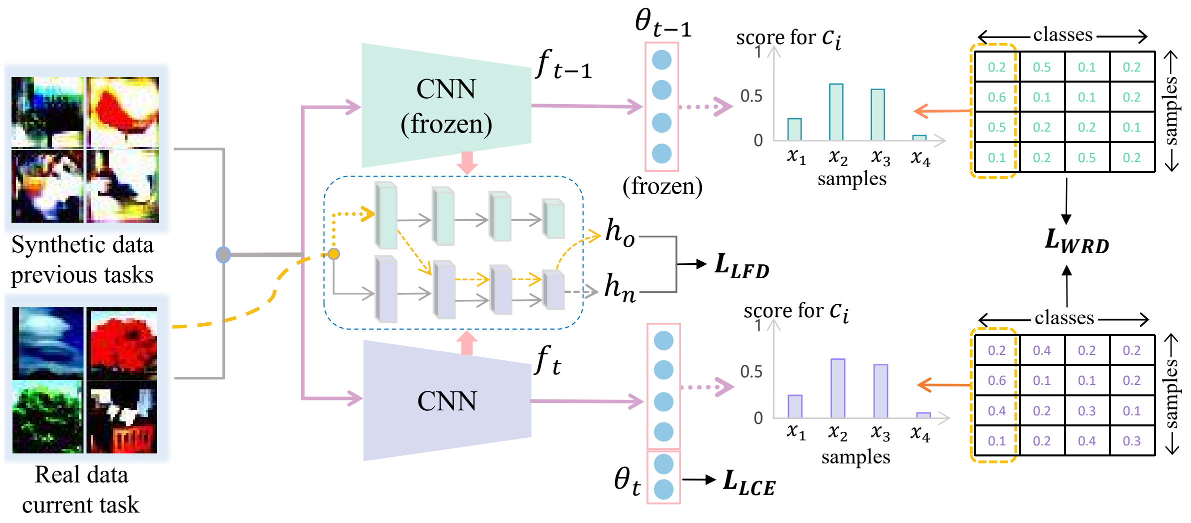

Fig. 1. Overview of our RFD. The model $\theta_t \circ f_t$ is being trained for the current task t using the localized CE loss L_{LCE}, the weak-constraint RD loss L_{WRD} and the low-level FD loss L_{LFD}. h denotes the extracted features, c_i denotes the i-th class.

M_t sample pairs $\{(x_i^t, y_i^t)\}_{i=1}^{M_t}$ belonging to classes ζ_t (ζ_t represents the classes set of t, $\zeta_t \cap \zeta_{t+1} = \emptyset$), where x_i^t is the image sample and y_i^t is the corresponding ground-truth label. The goal of CIL is to sequentially train a model that can eventually classify all task data without task identifiers.

In the CIL classification tasks, the architecture of deep neural networks typically comprises two components: a feature extractor, denoted as f_t, and a linear classifier, denoted as θ_t. First, the feature extractor f_t is to transform the input data x into a deeply embedded feature space, denoted as $h = f_t(x) \in R^d$. Then, the linear classifier θ_t is to perform the classification of the embedded feature h and outputs the logits $z = \theta_t(h) \in R^{|\zeta_{[1,t]}|}$, where $\zeta_{[1,t]}$ denotes the set of all classes from the task 1 to the task t. When a new task $t+1$ emerges, the existing classifier θ_t is augmented by integrating an additional classification head corresponding to the number of new classes (i.e., $|\zeta_{t+1}|$), thereby evolving into θ_{t+1}. Distinctively, the DFCIL differs from the CIL in that the data of previous tasks cannot be accessed during training.

B. Our RFD Method

The framework of our RFD is illustrated in Fig. 1. We present the RFD through the following two distinct phases: representation learning and classifier refinement.

1) Representation Learning: The primary objective of the representation learning phase is to integrate and assimilate representations of both old and new knowledge within the feature space. Toward this objective, three distinct components are applied to facilitate a more comprehensive and robust representation of old and new knowledge, as delineated below.

Localized Cross-entropy Function As done in previous works [19], [20], we adopt the model inversion technique to train an image synthesizer upon the completion of each learning task, which is employed to synthesize the old data replacing the real old data of previous tasks during training new tasks. To prevent the issue of domain shift caused by the utilization of synthetic data (the detailed information can refer to [19], [20]), we exclusively utilize the cross-entropy (CE) function on the new data of current task for classification at this phase. We denote this as the localized CE function. Given the data of current task (x_{new}, y_{new}), the softmax function $\mathrm{sf}(\cdot)$, the localized CE loss L_{LCE} is formulated as:

$$L_{LCE} = \mathrm{CrossEntropy}(\mathrm{sf}(\theta_t(f_t(x_{new}))), y_{new}), \quad (1)$$

Weak-constraint Relation Distillation Strategy With the arrival of new tasks, the retention of old knowledge becomes critical. To retain old knowledge while minimizing damage to new task learning, we introduce the weak-constraint relation distillation strategy which focuses on the correlation relationships among predicted probabilities of multiple image samples within a class. Concretely, we assume that p^o and p^n are the probability distributions predicted by the old and new model on K classes respectively. The Pearson correlation coefficient is employed as an relaxing metric to quantify the degree of similarity between these two probability distributions, which is mathematically formulated as follows:

$$\rho(p^o, p^n) = \frac{\sum_{i=1}^{K}(p_i^o - \bar{p^o})(p_i^n - \bar{p^n})}{\sqrt{\sum_{i=1}^{K}(p_i^o - \bar{p^o})^2}\sqrt{\sum_{i=1}^{K}(p_i^n - \bar{p^n})^2}}, \quad (2)$$

where p_i^o (or p_i^n) denotes the i-th ($1 \le i \le K$) element of p^o (or p^n), $\bar{p^o}$ and $\bar{p^n}$ denote the respective means of p^o and p^n.

For the current task t, let z^t (or z^{t-1}) denote the output logits of the current model (or old model, the task $t-1$) and Y^t (or Y^{t-1}) denote the predicted probabilities of the current model (or old model). Given the input data x, the logits z^t are obtained by: $z^t = \theta_t(h_t) = \theta_t(f_t(x))$. Let $z^t[, : t-1]$ denote the logits part corresponding to old tasks. As the calculation at this juncture is confined to the values associated with the old tasks, the predicted probabilities for these tasks, denoted as

$\tilde{Y^t} \in R^{B\times\left|\zeta_{[1,t-1]}\right|}$ (B represents the batchsize), are calculated based on $z^t[,:t-1]$, utilizing the formula:

$$\tilde{Y^t} = \mathrm{softmax}(z^t[,:t-1]/\tau), \tag{3}$$

where τ serves as a crucial temperature parameter to regulate the softness of $z^t[,:t-1]$. Similarity, the logits of old model z^{t-1} are obtained by: $z^{t-1} = \theta_{t-1}(h_{t-1}) = \theta_{t-1}(f_{t-1}(x))$. The predicted probabilities of old model $Y^{t-1} \in R^{B\times\left|\zeta_{[1,t-1]}\right|}$ are computed as:

$$Y^{t-1} = \mathrm{softmax}(z^{t-1}/\tau). \tag{4}$$

Thus, the weak-constraint RD loss L_{WRD} can be calculated using the following formula:

$$L_{WRD} = \frac{1}{\left|\zeta_{[1,t-1]}\right|}\sum_{i=1}^{\left|\zeta_{[1,t-1]}\right|}(1-\rho(Y^{t-1}_{:,i}, \tilde{Y^{t}_{:,i}})), \tag{5}$$

where $Y_{:,i}$ refers to the probability scores of all samples in a given batch corresponding to the class c_i, and $c_i \in \zeta_{[1,t-1]}$.

Note that the weak-constraint RD is applied to the synthesized old data and current new data separately, rather than in a collective manner. This deliberate design is adopted to avoid the issue where the learning weights are biased towards new data due to the imbalanced class distribution and limited volume of synthesized old data, thereby impeding the retention of previously learned old knowledge.

Low-level Feature Distillation Strategy Although the weak-constraint RD facilitates the retention of previously learned old knowledge, it's relaxed alignment between probability distributions could result in the loss of some fundamental general knowledge. To overcome this limitation, we further introduce the low-level feature distillation strategy to further enhance the model performance.

Concretely, we first input the current new data x_{new} into the feature extractor f_t of the current model, to obtain the current embedded features $h_n \in R^{B\times C\times W\times H}$ (here C refers to the channel), which is mathematically represented as: $h_n = f_t(x_{new})$. Meanwhile, we input x_{new} into the (frozen) feature extractor f_{t-1} of the old model, the first-stage features h_o^1 are extracted by $h_o^1 \leftarrow f_{t-1}(x_{new})$. Then the extracted features h_o^1 are fed into the feature extractor f_t of the current model to obtain the old embedded features $h_o \in R^{B\times C\times W\times H}$, which is represented as: $h_o = f_t'(h_o^1)$, where f_t' denotes the convolutional layers of the latter three stages of f_t. Finally, we compute the mean values across the spatial dimension ($W\times H$) and get the two channel-wise embedded features $\hat{h}_n \in R^{B\times C}$ and $\hat{h}_o \in R^{B\times C}$, respectively. Thus, the low-level FD loss L_{LFD} is formulated as:

$$L_{LFD} = \left\|\hat{h}_o - \hat{h}_n\right\|_2^2. \tag{6}$$

2) Classifier Refinement: To separate the decision boundaries between the old and new classes and make a better classification, we freeze the feature extractor component of the current model, opting solely to finetune the classifier. Concretely, we utilize the global CE function to classify both the synthesized old data and the current new data. Additionally, the weak-constraint RD strategy is employed to ensure the stability of decision boundaries in previous tasks by preserving the relative positions of samples within each class. Given the current new data (x_{new}, y_{new}) and the synthesized old data (x_{old}, y_{old}), the global CE loss L_{GCE} is defined as follows:

$$L_{GCE} = \mathrm{CrossEntropy}(\mathrm{sf}(\theta_t(f_t(x_{new}\cup x_{old}))), y_{new}\cup y_{old}). \tag{7}$$

Finally, the total loss L_{total} can be formulated as:

$$L_{total} = \lambda_{ce}L_{CE} + \lambda_{kd}L_{WRD} + \lambda_{kd}L_{LFD}, \tag{8}$$

where λ_{ce} and λ_{kd} are the weight parameters. The L_{CE} refers to either L_{LCE} or L_{GCE}, corresponding to the two respective phases. And in the second phase, the L_{LFD} is not utilized.

III. Experiments

A. Experimental Setup

Datasets The performance of the model is evaluated using five standard benchmark datasets: S-CIFAR-10 [24], S-CIFAR-100 [24], S-Tiny-ImageNet [25], S-ImageNet100 [26], and S-ImageNet200-R [27]. These datasets are equally split into 5, 10, and 20 tasks, with the exception of S-CIFAR-10, which is split into 5 tasks. In accordance with the training protocol delineated in [19], [20], for the first three datasets, three random orders of classes (produced by setting seeds as 0,1,2) are utilized for sequential training across three independent runs. For the latter two datasets, a random order of classes is utilized for sequential training once.

Implementation Details We implement our RFD based on the framework of R-DFCIL [20]. For the CIFAR datasets, we adopt a modified 32-layer ResNet [28] (without pretraining) as the backbone. For other ImageNet datasets, we adopt the ResNet18 [28] (without pretraining) as the backbone. We train the model with Stochastic Gradient Descent (SGD) optimizer with an initial learning rate $\eta = 0.1$. The temperature parameter τ is set to 4 in all experiments. The weight parameters λ_{ce} and λ_{kd} are determined through a grid search for each dataset. The batchsizes for the CIFAR and ImageNet datasets are set to 128 and 64, respectively. For more experimental details, please refer to [20]. Following [19], [20], the two metrics including the last incremental accuracy Acc_N (Acc) and average incremental accuracy $\bar{Acc}_N$ ($\bar{Acc}$) are adopted to evaluate the model performance. It is worth noting that for both metrics, higher values are indicative of superior model performance. The code is available at link .

B. Main Results

We compare our RFD with the following state-of-the-art methods: LwF [22], DeepInversion [15], ABD [19], FeTrIL [23] and R-DFCIL [20]. The comparative results are shown in Table I. We can see that: **(1)** Our RFD consistently achieves the highest accuracies across all three datasets, outperforming the state-of-the-arts in various sequential tasks. These results demonstrate the effectiveness of our proposed RFD in both short and long sequential tasks under the DFCIL

TABLE I
COMPARISON TO THE STATE-OF-THE-ARTS ON THE S-CIFAR-10, S-CIFAR-100 AND S-TINY-IMAGENET DATASETS. ALL DFCIL METHODS ARE TRAINED FROM SCRATCH. THE ACCURACY RESULTS ARE REPORTED ACROSS THREE INDEPENDENT RUNS.

	Method	S-CIFAR-10	S-CIFAR-100			S-Tiny-ImageNet		
	Tasks	$N = 5$	$N = 5$	$N = 10$	$N = 20$	$N = 5$	$N = 10$	$N = 20$
	Upper Bound	78.99 (±0.16)	70.67 (±0.16)	70.67 (±0.16)	70.67 (±0.16)	55.39 (±0.33)	55.39 (±0.33)	55.39 (±0.33)
Acc (↑)	LwF [22]	19.89 (±0.07)	17.00 (±0.10)	9.20 (±0.00)	4.70 (±0.10)	14.45 (±0.48)	8.18 (±0.32)	4.51 (±0.12)
	DeepInversion [15]	20.34 (±0.22)	18.80 (±0.30)	10.90 (±0.60)	5.70 (±0.30)	14.73 (±0.62)	8.75 (±0.45)	5.15 (±0.25)
	ABD [19]	55.38 (±4.29)	47.36 (±0.48)	36.19 (±0.93)	22.29 (±0.65)	30.56 (±0.22)	22.87 (±0.67)	15.20 (±1.01)
	FeTrIL [23]	56.23 (±3.24)	47.66 (±0.31)	41.28 (±0.64)	32.52 (±0.29)	35.46 (±0.46)	30.31 (±0.03)	25.57 (±0.17)
	R-DFCIL [20]	57.56 (±4.12)	50.47 (±0.43)	42.37 (±0.72)	30.75 (±0.12)	35.89 (±0.75)	29.58 (±0.51)	24.43 (±0.82)
	RFD (Ours)	**59.11 (±3.18)**	**53.01 (±0.23)**	**45.16 (±0.72)**	**33.41 (±0.23)**	**37.69 (±0.50)**	**32.64 (±0.27)**	**25.63 (±0.64)**
$\bar{Acc}$ (↑)	ABD [19]	70.38 (±1.86)	63.23 (±1.49)	56.61 (±1.93)	45.10 (±2.01)	45.30 (±0.50)	41.05 (±0.54)	34.74 (±0.91)
	FeTrIL [23]	70.66 (±1.05)	61.55 (±0.21)	55.09 (±1.35)	45.70 (±1.49)	46.95 (±0.27)	43.17 (±0.29)	39.03 (±0.72)
	R-DFCIL [20]	73.81 (±1.56)	64.85 (±1.78)	59.41 (±1.76)	48.47 (±1.90)	48.96 (±0.40)	44.36 (±0.18)	39.34 (±0.18)
	RFD (Ours)	**74.91 (±1.23)**	**66.24 (±1.38)**	**60.87 (±1.44)**	**50.45 (±0.86)**	**49.30 (±0.74)**	**45.67 (±0.19)**	**39.36 (±0.44)**

TABLE II
COMPARISON TO THE STATE-OF-THE-ART R-DFCIL [20] ON THE S-IMAGENET100 AND S-IMAGENET200-R DATASETS.

	Method	S-ImageNet100			S-ImageNet200-R		
	Tasks	$N = 5$	$N = 10$	$N = 20$	$N = 5$	$N = 10$	$N = 20$
	Upper Bound	77.46	77.46	77.46	44.47	44.47	44.47
Acc (↑)	R-DFCIL	53.10	42.28	30.28	18.75	12.50	7.23
	RFD (Ours)	**56.54**	**44.04**	**31.52**	**22.05**	**16.07**	**10.25**
$\bar{Acc}$ (↑)	R-DFCIL	68.15	59.10	47.33	26.96	20.50	14.80
	RFD (Ours)	**70.14**	**60.54**	**48.64**	**29.37**	**24.65**	**17.91**

TABLE III
ABLATIVE RESULTS OF OUR RFD ON S-CIFAR-100 ($N = 10$) AND S-TINY-IMAGENET ($N = 10$). 'WRD' REFERS TO WEAK-CONSTRAINT RD STRATEGY, 'LFD' REFERS TO LOW-LEVEL FD STRATEGY.

Method	S-CIFAR-100		S-Tiny-ImageNet	
	Acc (↑)	$\bar{Acc}$ (↑)	Acc (↑)	$\bar{Acc}$ (↑)
Base	9.22 (±0.10)	28.03 (±0.34)	7.93 (±0.09)	23.70 (±0.40)
Base+WRD	44.66 (±0.79)	60.47 (±1.60)	31.98 (±0.48)	45.51 (±0.15)
Base+WRD+LFD	**45.16 (±0.72)**	**60.87 (±1.44)**	**32.64 (±0.27)**	**45.67 (±0.19)**

setting. **(2)** Our RFD exhibits a significant superiority over the second best R-DFCIL. This is evidenced by the fact that the RFD surpasses R-DFCIL by an average value 2.64/1.61 ($Acc/\bar{Acc}$) on the S-CIFAR-100, and by an average value 2.02/0.56 on the S-Tiny-ImageNet. These accuracy differences clearly demonstrate that our RFD effectively retains old knowledge while enhancing the capacity for acquiring new knowledge, thus significantly improving overall model performance under the DFCIL setting.

Further, we present the comparative results of our RFD and the second-best R-DFCIL across two larger and challenging datasets, as shown in Table II. We can see that our RFD achieves a superior performance over the R-DFCIL, evidenced by an average enhancement of 2.15/1.58 ($Acc/\bar{Acc}$) on the S-ImageNet100, and a 3.30/3.22 average increment on the S-ImageNet200-R. These empirical results, spanning five distinct datasets, fully validate the effectiveness of our proposed RFD, demonstrating its robust applicability across a diverse range of datasets, from small to large scales.

C. Ablation Study

To demonstrate the impact of each proposed component (the weak-constraint RD & the low-level FD) on the performance of our RFD, we conduct the ablation study on S-CIFAR-100 ($N = 10$) and S-Tiny-ImageNet ($N = 10$). We take the original framework only with the CE loss function as the baseline (denoted as Base). On the basis of Base, we first add the weak-constraint RD loss function to the framework, which is denoted as Base+WRD. Then, we add the low-level FD loss function to the framework, which is denoted as Base+WRD+LFD, i.e., consisting of our full RFD.

The ablative results are shown in Table III. It can be observed that: **(1)** When only using the CE loss, the model exhibits lower accuracy on both datasets. **(2)** The application of the weak-constraint RD results in a substantial increase in accuracy for both datasets. This indicates that the weak-constraint RD enables the model to retain a large amount of useful old knowledge while acquiring more new knowledge, and the old and new knowledge are well integrated. **(3)** The application of the low-level FD (Base+WRD+LFD) further improves the accuracy, and yields the best results. This indicates that the low-level FD enforces the model to retain more fundamental general knowledge. Such knowledge is instrumental in preserving old knowledge and assimilating new knowledge, thereby culminating in optimal model performance. Overall, these results fully demonstrate each proposed component (i.e., the weak-constraint RD or the low-level FD) contributes to the model performance. Particularly, the weak-constraint RD plays an extremely important role in our method.

IV. CONCLUSION

This paper presents a simple yet effective method called RFD for DFCIL to retain more old knowledge while facilitating new knowledge acquisition, thus enhancing the model performance. Concretely, our approach first introduces a weak-constraint relation distillation strategy to retain more useful old knowledge and promote the assimilation of new knowledge by learning intra-class sample relationships. Further, to enhance the preservation of old knowledge and to refine the integration of new knowledge, we introduce a low-level feature distillation strategy to dig deep and retain the fundamental general knowledge. Extensive experiments on five benchmarks demonstrate that our method significantly outperforms the state-of-the-arts under the DFCIL setting.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (62376274, 62437002).

References

[1] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[2] J. Park, M. Kang, and B. Han, "Class-incremental learning for action recognition in videos," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 698–13 707.

[3] M. Kang, J. Park, and B. Han, "Class-incremental learning by knowledge distillation with adaptive feature consolidation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 071–16 080.

[4] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. Van De Weijer, "Class-incremental learning: survey and performance evaluation on image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5513–5533, 2022.

[5] D.-W. Zhou, Q.-W. Wang, Z.-H. Qi, H.-J. Ye, D.-C. Zhan, and Z. Liu, "Class-incremental learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[6] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of Learning and Motivation*. Elsevier, 1989, vol. 24, pp. 109–165.

[7] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[8] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu, "Prototype augmentation and self-supervision for incremental learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5871–5880.

[10] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 139–149.

[11] Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy *et al.*, "Dualprompt: Complementary prompting for rehearsal-free continual learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 631–648.

[12] J. S. Smith, L. Karlinsky, V. Gutta, P. Cascante-Bonilla, D. Kim, A. Arbelle, R. Panda, R. Feris, and Z. Kira, "Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 909–11 919.

[13] Z. Qiu, Y. Xu, F. Meng, H. Li, L. Xu, and Q. Wu, "Dual-consistency model inversion for non-exemplar class incremental learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 025–24 035.

[14] Z. Meng, J. Zhang, C. Yang, Z. Zhan, P. Zhao, and Y. Wang, "Diffclass: Diffusion-based class incremental learning," in *European Conference on Computer Vision*. Springer, 2025, pp. 142–159.

[15] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8715–8724.

[16] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[18] S. Li, T. Su, X. Zhang, and Z. Wang, "Continual learning with knowledge distillation: A survey," *Authorea Preprints*, 2024.

[19] J. Smith, Y.-C. Hsu, J. Balloch, Y. Shen, H. Jin, and Z. Kira, "Always be dreaming: A new approach for data-free class-incremental learning," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9374–9384.

[20] Q. Gao, C. Zhao, B. Ghanem, and J. Zhang, "R-dfcil: Relation-guided representation learning for data-free class incremental learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 423–439.

[21] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[22] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[23] G. Petit, A. Popescu, H. Schindler, D. Picard, and B. Delezoide, "Fetril: Feature translation for exemplar-free class-incremental learning," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3911–3920.

[24] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Handbook of Systemic Autoimmune Diseases*, vol. 1, no. 4, 2009.

[25] A. Banerjee and V. Iyer, "Cs231n project report-tiny imagenet challenge," 2015.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[27] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.