DiAPR: Dimensionally-Allocated Prototype Refinement for Non-Exemplar Class Incremental Learning

Ruixuan Gao[1], **Qijun Zhao**[1*], **Keren Fu**[1]

[1]College of Computer Science, Sichuan University
grx@stu.scu.edu.cn, {qjzhao, fkrsuper}@scu.edu.cn

Abstract

Non-Exemplar Class Incremental Learning (NECIL) strives to preserve classification performance in an evolving data stream without revisiting old-class exemplars. Current methods mitigate catastrophic forgetting by replaying and augmenting historical prototypes as surrogates for old classes. However, they treat prototypes as holistic representations for global-level augmentations, which overlook dimensional semantic disparity and old-new class relationships, failing to maintain old-class discriminability and adaptability to the evolving feature space. To address this challenge, we propose **Di**mensionally-**A**llocated **P**rototype **R**efinement (DiAPR), a granular framework that progressively refines prototypes to exhibit class separability in the new feature space through three modules. Specifically, Distribution-aware Pairing (DAP) captures old-new class semantic consistency to guide Granular Semantic Allocation (GSA) in dimension-wise conflation, while Cross-Dimensional Transition (CDT) enhances cross-dimensional dependencies. The resulting prototypes sharpen classifier decision boundaries. Moreover, CDT inherently enables softened feature alignment, thereby yielding a more compatible feature space. Extensive experiments demonstrate DiAPR's superiority, with improvements over SOTA by 2.35%, 0.70%, 0.96% on three CIFAR-100 settings, 1.03%, 0.54%, 0.40% on Tiny-ImageNet, and 0.60% on ImageNet-Subset.

Introduction

Modern deep learning models achieve human-competitive classification performance when trained on full labeled datasets (Lin et al. 2014; Deng et al. 2009). However, in practical settings where data arrives sequentially, these models suffer from catastrophic forgetting, wherein learning new classes degrades performance on previously learned ones. To address this issue, Incremental Learning (IL) (Douillard et al. 2022; Yan, Xie, and He 2021; Wang et al. 2022) has been proposed to acquire new knowledge while preserving old knowledge.

Most IL methods rely on exemplar rehearsal, which preserves prior knowledge by storing subsets of old data and retraining models jointly with new data. Despite their effectiveness, such approaches encounter two limitations: the unavailability of previous data due to privacy sensitivity (*e.g.*, medical imaging), and resource-constrained environments due to storage overhead. These constraints motivate Non-Exemplar Class Incremental Learning (NECIL), a challenging setting in which models must learn continuously without access to any prior exemplars.

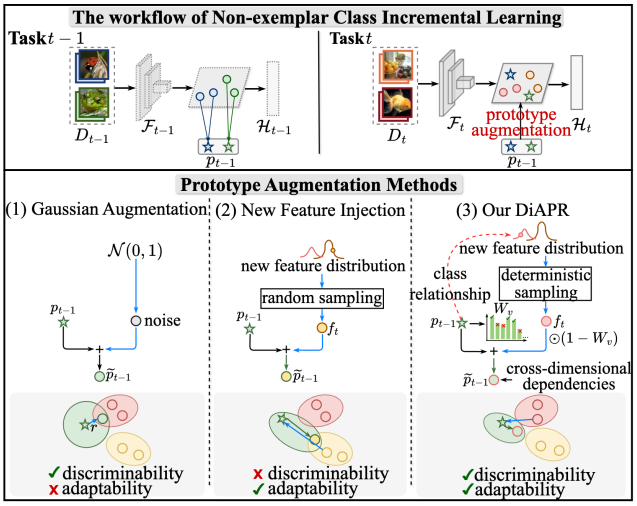


Figure 1: Comparative Analysis of Prototype Augmentation-based NECIL Methods: (1) Gaussian Noise Augmentation perform holistic perturbations on prototypes to expand outdated manifolds but fail to adapt to new feature space. (2) New Feature Injection indiscriminately merges new patterns into historical prototypes, compromising discriminative semantics. (3) Our method introduces dimension-wise prototype refinement, balancing semantic discriminability and locational adaptability, sharpening decision boundaries.

Compared to dynamically expanding model architectures (Sun et al. 2023; Roy et al. 2023) or employing generative models (Liu et al. 2024) for knowledge preservation, prototype-based NECIL methods achieve knowledge preservation through replaying historical class mean representations (*i.e.*, prototypes), which require neither excessive storage capacity nor additional model training. Yet the sparsity of historical prototypes (typically just one per class) induces decision boundary skew when new tasks are introduced. Re-

*Corresponding author.

cent works have developed prototype augmentation strategies to enhance prototype's density. Among these methods, a subset (Zhu et al. 2021b; Shi et al. 2023) enriches historical prototypes through repeated sampling or adding Gaussian noise. However, as the feature space evolves, historical prototypes inevitably shift and become outdated, while their uniformly perturbed variants remain confined to obsolete distributions, failing to promote effective adaptation (Fig. 1(1)). Other methods (Shi and Ye 2023; Zhai et al. 2024) augment historical prototypes by randomly injecting new features to enhance adaptability. However, their indiscriminate fusion across all dimensions dilutes old class-specific information and random sampling of new features tends to generate outliers (Fig. 1(2)). The above global-level prototype augmentation methods overlook dimensional importance disparity within prototypes and the old-new class relationship. As a result, they struggle to balance the plasticity required to adapt to the new feature space and the stability needed to preserve critical semantics of old classes, ultimately blurring the decision boundaries between old and new classes.

Feature dimensions differ in their importance for discriminative class separation. Dimensions with high variance typically encode stronger class-specific semantics, while those with low variance tend to be less informative. As shown in Fig. 2, retaining only higher-variance dimensions preserves a distribution similar to the original, whereas retaining only lower-variance dimensions results in scattered distributions and blurred decision boundaries. *This suggests that fine-grained preservation of higher-variance informative dimensions while repurposing less critical ones could serve as an effective strategy for adapting to evolving feature spaces without disrupting previously learned information.*

To this end, we propose Dimensionally-Allocated Prototype Refinement (DiAPR), a fine-grained framework that preserves critical semantic information in historical prototypes while ensuring their appropriate location in the evolving feature space, thereby sharpening decision boundaries (Fig. 1(3)). DiAPR progressively refines historical prototypes through three modules. Specifically, a Distribution-Aware Pairing (DAP) module forms semantically consistent pairs between historical prototypes and new features. These pairs reflect old-new class relationships and avoid outlier generation when fed into a Granular Semantic Allocation (GSA) for dimension-wise conflation. Higher-variance dimensions are minimally adjusted to retain old-class critical semantics, while others are allocated to conflate paired new features, enabling prototypes to locate appropriately in the evolving feature space. Given GSA's dimension-independent allocation, the Cross-Dimensional Transition (CDT) further captures cross-dimensional dependencies during task transitions in closed form. This enables prototypes to better integrate semantics across dimensions in new tasks, enhancing their robustness.

While capturing cross-dimensional dependencies, the CDT naturally enables softened feature alignment, fostering a flexible feature space. Traditional Knowledge Distillation (KD) rigidly enforces old-new feature alignment and stifles semantic evolution by treating natural semantic growth as

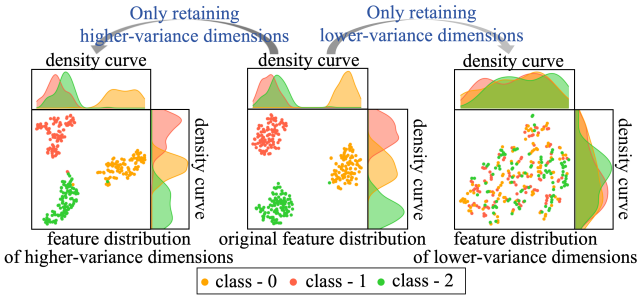


Figure 2: Visualization of feature distributions when retaining different dimensions. Higher-variance dimensions (left) preserve class separability, while lower-variance dimensions (right) exhibit overlapping distributions.

unwanted deviations. To compensate for this rigidity, CDT pre-calibrates old features to the new task. This collaboration avoids over-regularization, yielding a more compatible feature space that, together with sharpened classifier boundaries, enhances NECIL performance.

Our approach consistently achieves state-of-the-art performance across CIFAR-100, Tiny-ImageNet, and ImageNet-Subset benchmarks. In summary, the contributions of this paper include: (1) A granular dimension-wise refinement strategy that balances prototype adaptation to new feature spaces and preservation of old discriminative semantics, thereby sharpening classifier decision boundaries. (2) A softened feature alignment mechanism that fosters a compatible feature space for the feature extractor. (3) Extensive experiments validate that our DiAPR framework outperforms existing state-of-the-art methods across various datasets and experimental settings.

Related Work

Incremental Learning

Incremental Learning (IL) aims to accumulate knowledge continuously through training on a sequence of tasks. Some works (Rebuffi et al. 2017; Chaudhry et al. 2018) rehearse exemplars from previous tasks during new task training. Regularization-based methods (Riemer et al. 2018; Rajasegaran et al. 2020; Tang et al. 2021) impose constraints on the novel model to maintain performance on previous exemplars that is highly similar to the old model's. Dynamic architecture methods (Schwarz et al. 2018; Liu, Schiele, and Sun 2021; Zhou et al. 2022; Douillard et al. 2022; Gao et al. 2023) enhance model adaptability by introducing additional task-specific learnable parameters, albeit at the cost of increased computational and memory demands. All the above methods assume access to stored exemplars, which raises concerns under strict privacy constraints and limited memory capacity.

Non-exemplar Class Incremental Learning

Considering data privacy and limited storage capability, researchers have increasingly focused on Non-Exemplar Class Incremental Learning (NECIL), which seeks to retain old

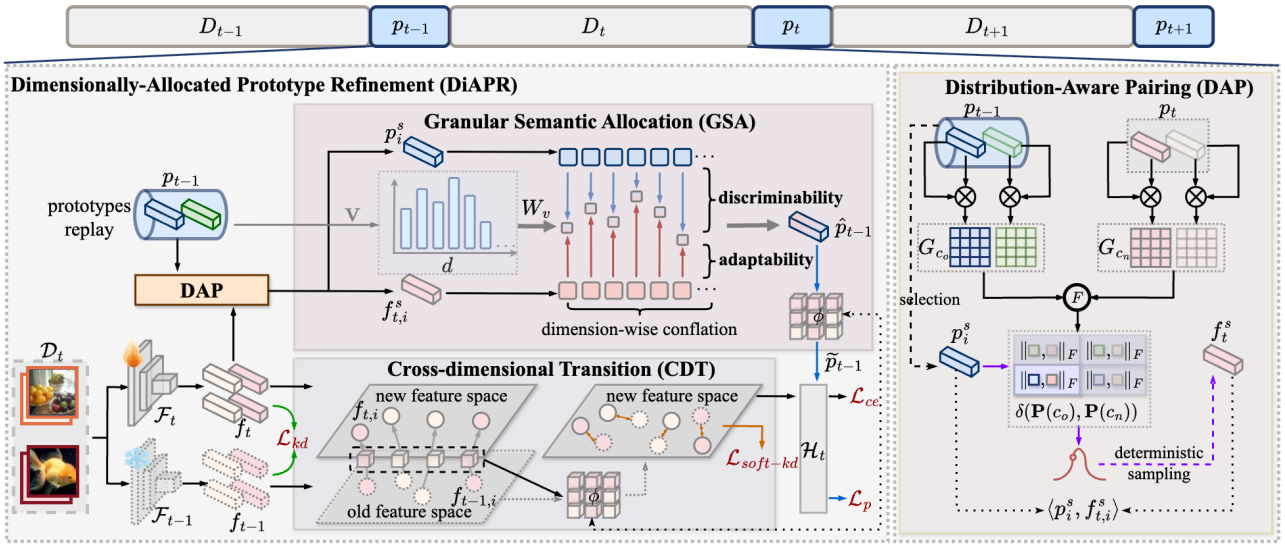


Figure 3: Overview of the DiAPR framework. During task t, historical prototypes p_{t-1} and the new task D_t collaboratively support the stability-plasticity trade-off through three novel modules. DAP forms semantically consistent pairs by approximating class distribution discrepancies. These pairs are then fed into GSA for dimension-wise conflation via normed variance. Finally, CDT further enhances prototypes by explicitly constructing a closed-form old-to-new task transition with cross-dimensional dependencies. Meanwhile, CDT enables softened feature alignment through $\mathcal{L}_{soft-kd}$, yielding a more compatible feature space.

knowledge without storing exemplars from previous tasks. Some methods (Petit et al. 2023; Liu et al. 2024; Roy et al. 2023; Gomez-Villa et al. 2024; Sun et al. 2023) aim to strike a balance between learning new knowledge and retaining old knowledge through model expansion or synthetic sample generation. However, the incorporation of additional learnable modules increases memory overhead and deviates the model from its primary objective.

Common strategies (Yu et al. 2020; Pelosin et al. 2022; Zhai et al. 2024) employ Knowledge Distillation (KD) to constrain representation alignment between new and old models. However, rigid feature alignment indiscriminately mitigates all drifts, even the beneficial ones arising from new knowledge acquisition, thereby impairing model plasticity. Some methods (Gu, Shim, and Shkurti 2023; Li, Peng, and Zhou 2024) utilize learnable linear layers to protect old knowledge without compromising plasticity. Yet these learnable parameters rely on random initialization, which may lead to training instability.

Another popular line of work (Zhu et al. 2021b; Goswami et al. 2023) directly replays historical prototypes during novel task training. But compared to the abundance of new features, historical prototypes are relatively scarce, leading to classifier imbalance. Considering this issue, some methods (Shi et al. 2023; Dong et al. 2024) introduce Gaussian noise to expand the manifolds of stored prototypes. Other methods (Yu et al. 2020; Zhai et al. 2024) compute the difference between historical and new prototypes as an estimated drift to guide the expansion. PRAKA (Shi and Ye 2023) proposes randomly selecting new features and blending them with historical prototypes using fixed weights to further enrich their representation. However, these methods treat prototypes as holistic representations and perform global-level augmentation, failing to consider dimensional semantic importance and overlooking relationships between old and new classes. Consequently, they cannot simultaneously achieve the preservation of old critical semantics and adaptation to the evolving feature space.

Methodology

Preliminaries

NECIL aims to train a model that excels on both old and new tasks without access to historical data. Formally, given a sequence of tasks $\mathcal{D} = \{D_t\}_{t=0}^{T}$, where T is the total number of tasks. Each task $D_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{N_t}$ consists of N_t images $x_{t,i}$ and their corresponding labels $y_{t,i} \in C_t$, with C_t denoting the class space of the task t. These class spaces are non-overlapping, that is $C_i \bigcap C_j = \emptyset$, which allows for further evaluation of the model's generalization across different tasks. During the t-th task, the model comprises a feature extractor $\mathcal{F}_t$ and a classifier $\mathcal{H}_t$. These are updated incrementally as new tasks are introduced, building on the knowledge acquired from previous tasks. Here, $f_{t,i} = \mathcal{F}_t(x_{t,i}) \in \mathbb{R}^d$ is the feature of $x_{t,i}$, and $\mathcal{H}_t(f_{t,i}) \in \mathbb{R}^{\sum_{j=0}^{t}|C_j|}$ is the final prediction, where d means the feature dimension.

Overall Pipeline

The overall pipeline of DiAPR at task t is illustrated in Fig. 3. For a new task D_t, old features f_{t-1} and new features f_t are extracted from the frozen old feature extractor $\mathcal{F}_{t-1}$ and current feature extractor $\mathcal{F}_t$, respectively. Historical prototypes $p_{t-1} \in \mathbb{R}^{\sum_{i=0}^{t-1}|C_i| \times d}$ from prior tasks are progressively refined through three modules: **DAP** provides seman-

tically consistent paired for **GSA**'s dimension-wise discriminative conflation, while **CDT** enhances cross-dimensional dependencies. Refined prototypes $\widetilde{p}_{t-1}$ are fed into classifier $\mathcal{H}_t$ to sharpen decision boundaries. Additionally, CDT enables **softened feature alignment** via $\mathcal{L}_{soft-kd}$, fostering a more compatible feature space. Collectively, they boost NECIL's performance.

Distribution-Aware Pairing (DAP)

Semantically consistent pairs serve as a prerequisite to reflect the old-new class relationship, avoiding outlier generation and enabling historical prototypes to be more suitably located in the new feature space. Denote the class distribution of historical class $c_o \in \bigcup_{i=0}^{t-1} C_i$ and new class $c_n \in C_t$ as $\mathbf{P}(c_o)$ and $\mathbf{P}(c_n)$, respectively. Our goal is to estimate their distribution discrepancy $\delta(\mathbf{P}(c_o), \mathbf{P}(c_n))$ to guide semantically consistent pairing between historical prototypes and new features.

Given that historical prototypes p_{t-1} represented as class means offer incomplete descriptions of their underlying class distributions, we employ Gram matrices to capture second-order statistics for more comprehensive distributional approximations:

$$\mathbf{G}_{c_o} = (p_{t-1,c_o})^\top (p_{t-1,c_o}), \quad \mathbf{G}_{c_n} = (p_{t,c_n})^\top (p_{t,c_n}), \tag{1}$$

where $p_{t,c_n} = \frac{1}{N_{b,c_n}} \sum_{i=1}^{N_{b,c_n}} f_{t,i}$ is the new prototype, with N_{b,c_n} denoting the sample number of class c_n in the mini-batch per training iteration. $\mathbf{G}_{c_o}$ and $\mathbf{G}_{c_n}$ serve as statistical proxies for the true class distributions $\mathbf{P}(c_o)$ and $\mathbf{P}(c_n)$. Distribution discrepancy is quantified via the Frobenius norm:

$$\begin{aligned}\delta(\mathbf{P}(c_o), \mathbf{P}(c_n)) &= \|\mathbf{G}_{c_o} - \mathbf{G}_{c_n}\|_F \\ &= \sqrt{\sum_{(i,j)\in[d]^2} (\mathbf{G}_{c_o,(i,j)} - \mathbf{G}_{c_n,(i,j)})^2}.\end{aligned} \tag{2}$$

To construct pairs, we first select a historical prototype p_i^s from p_{t-1} with label $y_{p,i}^s$, then identify the most semantically consistent new class distribution by minimizing δ:

$$\hat{k} = \underset{k\in C_t}{\arg\min}\, \delta\left(\mathbf{P}(y_{p,i}^s), \mathbf{P}(k)\right), \tag{3}$$

where $\hat{k}$ is the index of the target new class distribution. We perform deterministic sampling to draw a new feature $f_{t,i}^s \sim \mathbf{P}(\hat{k})$, forming a pair $\langle p_i^s, f_{t,i}^s \rangle$. This pair construction process is repeated until the number of pairs equals N_b, *i.e.*, $\{\langle p_i^s, f_{t,i}^s \rangle\}_{i=1}^{N_b}$. These semantically consistent pairs guide historical prototypes to converge actively toward their matched new features during adaptation to the new feature space. Adhering to the principle that semantically similar classes are spatially proximal, this mechanism facilitates optimal prototype location in the new space.

Granular Semantic Allocation (GSA)

Global-level augmentations, which perform coarse-grained adjustments on historical prototypes, struggle to simultaneously preserve key old semantics and adapt to the new feature space. To address this, the GSA module conducts discriminative dimension-wise conflation for granular prototype refinement.

First, we quantify the contribution of each prototype dimension to class discriminability. As illustrated in Fig. 2, variance $\mathbf{V} \in \mathbb{R}^d$ serves as an effective metric:

$$\begin{aligned}&\mathbf{V} = \frac{1}{C_{t-1}^{all}} \sum_{i=1}^{C_{t-1}^{all}} (p_{t-1,i} - \mu_{t-1})^2, \\ &with \quad \mu_{t-1} = \frac{1}{C_{t-1}^{all}} \sum_{i=1}^{C_{t-1}^{all}} p_{t-1,i}, \quad C_{t-1}^{all} = \sum_{i=0}^{t-1} |\, C_i \,|,\end{aligned} \tag{4}$$

where higher variances indicate stronger class separability. To avoid the influence of extreme values, we apply log normalization to the variance and derive dimension-wise importance weights $W_v \in \mathbb{R}^d$:

$$W_v = \frac{log(\mathbf{1} + \mathbf{V}) - min(log(\mathbf{1} + \mathbf{V}))}{max(log(\mathbf{1} + \mathbf{V})) - min(log(\mathbf{1} + \mathbf{V}))}, \tag{5}$$

where $\mathbf{1} \in \mathbb{R}^d$ is an all-one vector, $min(\cdot)$ and $max(\cdot)$ denote taking the minimum and maximum of a vector's elements. For historical prototypes, higher-variance dimensions are minimally modified to retain their inherent discriminative semantics, ensuring stability of previous class representations. In contrast, lower-variance dimensions are allocated to absorb new features, participating in new task training to enhance plasticity. To modulate this process, we perform dimension-wise conflation on reliable semantically consistent pairs $\{\langle p_i^s, f_{t,i}^s \rangle\}_{i=1}^{N_b}$:

$$\hat{p}_{t-1} = \frac{1}{2}((1 + W_v) \odot p^s + (1 - W_v) \odot f_t^s), \tag{6}$$

where $\hat{p}_{t-1} \in \mathbb{R}^{N_b \times d}$ are the generated prototypes sharing the same label as y_p^s. When W_v approaches 1, the corresponding dimensions are prioritized to strengthen their inherent semantics, preserving old class discriminability. Conversely, dimensions with lower W_v are allocated to absorb new features, adapting to new tasks during training.

Cross-Dimensional Transition (CDT)

CDT serves two functions: (1) Considering GSA's dimension-independent operation, CDT models **cross-dimensional dependencies** during the transition from old to new task. These dependencies guide $\hat{p}_{t-1}$ to integrate semantics in new task training. (2) The transition process inherently performs old features pre-calibration, allowing **softened feature alignment** and thus fostering a more compatible feature space.

Cross-dimensional dependencies: Motivated by ridge regression's strengths in modeling cross-dimensional dependencies, we first model the transition from old to new tasks by formulating it as a ridge-regression problem, which aims to explore the optimal transition $\phi : \mathbb{R}^d \to \mathbb{R}^d$ that minimizes the cost between the transformed old features $\phi(f_{t-1})$ and the new features f_t. This process is formalized by the

Datasets	CIFAR-100						TinyImageNet						ImageNet-Subset	
Setting	5 tasks		10 tasks		20 tasks		5 tasks		10 tasks		20 tasks		10 tasks	
Methods	AVG	LAST	AVG	LAST	AVG	LAST	AVG	LAST	AVG	LAST	AVG	LAST	AVG	LAST
iCaRL-CNN*	51.07	40.12	48.66	39.65	44.43	35.47	34.64	22.31	31.15	21.10	27.90	20.46	50.53	41.08
iCaRL-NME*	58.56	49.74	54.19	45.13	50.51	40.68	45.86	33.45	43.29	33.75	38.04	28.89	60.79	51.90
EEIL*	60.37	52.35	56.05	47.67	52.34	41.59	47.12	34.24	45.01	34.26	40.50	30.14	63.34	54.19
LUCIR*	63.78	55.06	62.39	50.14	59.07	48.78	49.15	37.09	48.52	36.80	42.83	32.55	66.16	56.21
LwF_MC	45.93	36.17	27.43	15.47	20.10	15.88	29.12	17.12	23.10	12.33	17.43	8.75	31.18	20.01
MUC	49.42	38.45	30.19	19.57	21.27	15.65	32.58	17.98	26.61	14.54	21.95	12.70	35.07	22.65
PASS	63.47	55.67	61.84	49.03	58.09	48.48	49.55	41.58	47.29	39.28	42.07	32.78	61.80	50.44
IL2A	63.22	55.13	57.65	45.32	54.90	45.24	48.17	36.14	42.10	35.23	36.79	28.74	-	-
SSRE	65.88	56.33	65.04	55.01	61.70	50.47	50.39	41.67	48.93	39.89	48.17	39.76	67.69	57.51
FeTrIL	66.30	58.12	65.20	57.64	61.50	52.48	54.80	42.92	53.10	42.41	52.20	41.33	65.0	61.22
NAPA-VQ	70.44	-	69.04	-	67.42	-	52.77	-	51.78	-	49.51	-	68.83	-
MDPCR	64.18	-	60.87	-	57.37	-	47.42	-	46.89	-	43.94	-	66.20	55.00
PRAKA	70.02	61.55	68.86	60.41	65.86	56.20	53.32	46.36	52.61	45.16	49.83	40.58	68.98	61.30
DCMI	67.90	-	66.80	-	64.00	-	54.80	-	53.90	-	52.50	-	70.00	-
FCS	70.40	62.13	69.04	60.39	68.36	58.36	53.66	46.04	52.43	44.95	51.15	42.57	70.67	61.76
DS_AL	68.39	61.44	-	-	-	-	-	-	-	-	-	-	-	-
FGKSR†	68.17	59.02	70.13	57.90	66.86	54.25	54.88	44.97	52.72	43.35	51.68	41.94	70.18	61.42
DiAPR (Ours)	72.35	63.58	70.77	61.09	**69.62**	58.63	55.12	47.57	53.24	45.02	52.40	42.58	70.79	62.03
DiAPR (Ours)†	**72.79**	**64.58**	**70.83**	**62.17**	69.32	**58.64**	**55.91**	**48.28**	**54.44**	**45.90**	**52.90**	**43.45**	**71.27**	**62.80**

Table 1: Quantitative comparison of average accuracy (AVG) and final task accuracy (LAST) (%, higher is better) across different task settings on CIFAR-100, TinyImageNet, and ImageNet-Subset. Bold values indicate the best performance, underlined the second-best. Methods marked with * use rehearsal-based strategies. † denotes ViT-based implementations.

objective function: $\arg\min_{\phi}\|\phi(f_{t-1}) - f_t\|_2^2 + \lambda\|\phi\|_2^2$ (λ is a regularization parameter). Then the closed-form solution of this transition explicitly captures the embedded cross-dimensional dependencies:

$$\phi = (f_{t-1}^{\top} f_{t-1} + \lambda I)^{-1} f_{t-1}^{\top} f_t. \tag{7}$$

Unlike gradient-based optimization of learnable parameters (Experimental comparison can be found in Fig. 8), ϕ analytically models cross-dimensional dependencies through cross-covariance $f_{t-1}^{\top} f_t$ while preserving critical old semantic via auto-covariance $f_{t-1}^{\top} f_{t-1}$, enhancing historical prototypes in evolving feature space:

$$\widetilde{p}_{t-1} = \phi(\hat{p}_{t-1}). \tag{8}$$

The resulting prototypes $\widetilde{p}_{t-1}$ are ultimately used to train the classifier, encouraging sharper decision boundaries:

$$\mathcal{L}_p = L_{ce}(\mathcal{H}_t(\widetilde{p}_{t-1}), y_p^s), \tag{9}$$

where L_{ce} is the cross-entropy loss.

Softened feature alignment: Traditional KD loss directly aligns old and new features, thereby indiscriminately treating natural semantic growth as undesirable deviations. To compensate for this, CDT implicitly enables softened feature alignment by aligning transformed old features with new features during task transition construction. Specifically, the transition ϕ first adapts to semantic growth, thereby pre-calibrating old features to the new task. This process smooths benign drifts arising from semantic expansion, and subsequent feature alignment thus avoids excessive penalties that would hinder the learning of new knowledge:

$$\mathcal{L}_{soft-kd} = \|\phi(f_{t-1}) - f_t\|_2. \tag{10}$$

This fosters a more compatible feature space, thereby supporting both the retention of old knowledge and the acquisition of new knowledge.

Overall Loss Function

For task t, the total loss $\mathcal{L}_{all}$ includes $\mathcal{L}_{kd}$ and $\mathcal{L}_{soft\text{-}kd}$ for regularizing the feature extractor, $\mathcal{L}_{ce}$ and $\mathcal{L}_p$ for balancing the classifier:

$$\mathcal{L}_{all} = \underbrace{\mathcal{L}_{kd} + \alpha\mathcal{L}_{soft-kd}}_{\mathcal{F}_t} + \underbrace{\mathcal{L}_{ce} + \beta\mathcal{L}_p}_{\mathcal{H}_t}, \tag{11}$$

where $\mathcal{L}_{ce} = L_{ce}(\mathcal{H}_t(f_t), y_t)$ and $\mathcal{L}_{kd} = \|f_{t-1} - f_t\|_2$, follow standard configurations in NECIL, α and β are the hyperparameters that balance the respective components.

Experiments

Experimental Setting

Datasets: We evaluate our proposed DiAPR on three standard NECIL benchmarks: (1) **CIFAR-100** (Krizhevsky, Hinton et al. 2009) contains 100 classes with 600 images per class (500 training, 100 validation). We adopt three splitting protocols: 50 initial classes + 5/10 incremental tasks; 40 initial classes + 20 incremental tasks. (2) **TinyImageNet** (Le and Yang 2015) comprises 200 classes (500 training, 50 validation per class). We use 100 initial classes followed by 5, 10, or 20 tasks. (3) **ImageNet-Subset** is a curated subset of ImageNet (Deng et al. 2009; Russakovsky et al. 2015) with 100 classes (1,300 training, 50 validation per class), split into 50 initial classes + 10 incremental tasks.

Evaluation Metrics: We evaluate the performance of our DiAPR using three key metrics: (1) **Average Top-1 Accuracy (AVG)** represents the model's average performance on all prior learned tasks after completing the t-th task: $\text{AVG} = \frac{1}{T}\sum_{t=1}^{T} \text{acc}_t$, where acc_t denotes the average accuracy of all learned classes after training up to task

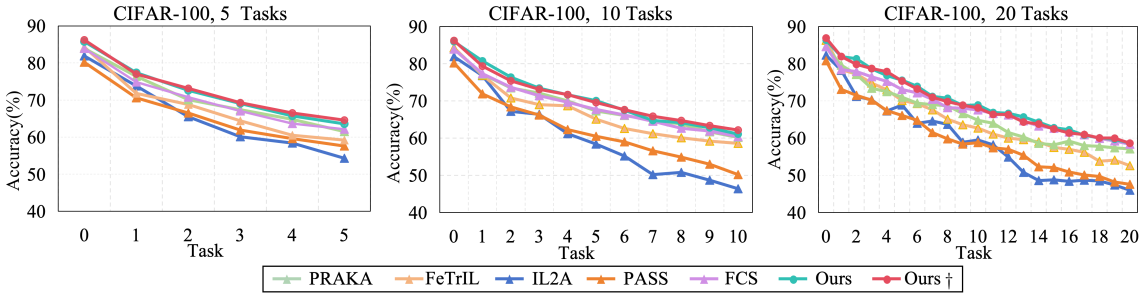


Figure 4: Accuracy curves of different methods across tasks on CIFAR-100. † denotes ViT-based implementation.

Datasets	CIFAR-100			TinyImageNet		
Methods	5 tasks	10 tasks	20 tasks	5 tasks	10 tasks	20 tasks
LwF_MC	44.23	50.47	55.46	54.26	54.37	63.54
MUC	40.28	47.56	52.65	51.46	50.21	58.00
PASS	25.20	30.25	30.61	18.04	23.11	30.55
IL2A	28.54	39.29	41.27	25.43	28.32	35.46
SSRE	18.37	19.48	19.00	9.17	14.06	14.20
FCS	12.20	16.70	15.90	-	-	-
FGKSR†	-	-	-	11.45	12.21	12.82
DiAPR†	**10.53**	**12.88**	**11.18**	**6.62**	**10.67**	**11.01**

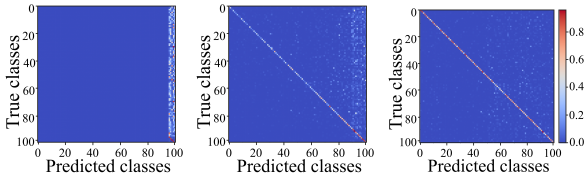
Table 2: Comparison of Average Forgetting (AF) (%, lower is better) on CIFAR-100 and TinyImageNet.

t. (2) **Final Task Accuracy (LAST)** reports the accuracy acc_T on the last learned task T. (3) **Average Forgetting (AF)** measures the average accuracy drop across all old tasks compared to their historical peak performance: $AF = \frac{1}{T-1}\sum_{t=0}^{T-1} \underset{i\in[t,T]}{max}(acc_{i,t} - acc_{T,t})$, where $acc_{i,t}$ means task i's accuracy after training up to task t.

Implementation Details: For a comprehensive evaluation, we adopt Vision Transformer (ViT, denotes †) (Dosovitskiy et al. 2020) and ResNet-18 (He et al. 2016) as feature extractors, respectively. The ViT is configured with 6 transformer blocks, a feature dimension of 384, and 12 self-attention heads. Training employs the Adam optimizer with a learning rate of 0.001, weight decay of $2e-4$, and a batch size of 64. Each task is trained for 300 epochs to ensure convergence. Hyperparameters for loss balancing are set as $\alpha = 1$ and $\beta = 10$. Experiments are implemented in PyTorch and conducted on an NVIDIA RTX 4090 GPU.

Quantitative Results

We compare our DiAPR method with state-of-the-art non-exemplar approaches: LwF (Li and Hoiem 2017), MUC (Liu et al. 2020), PASS (Zhu et al. 2021b), IL2A (Zhu et al. 2021a), SSRE (Zhu et al. 2022), FeTrIL (Petit et al. 2023), NAPA-VQ (Malepathirana, Senanayake, and Halgamuge 2023), MDPCR (Shi et al. 2023), PRAKA (Shi et al. 2023), FCS (Li, Peng, and Zhou 2024), FGKSR (Zhai et al. 2024), DS_AL (Zhuang et al. 2024) and DCMI (Qiu et al. 2024). For broader context, we also include rehearsal-based methods (iCaRL (Rebuffi et al. 2017), EEIL (Castro et al. 2018), LUCIR (Hou et al. 2019)) that store 20 exemplars per class.



(a) Finetuning (b) iCaRL (c) DiAPR† (Ours)

Figure 5: Normalized confusion matrices of different methods on CIFAR-100 under the 10-task setting.

Accuracy Analysis: Tab. 1 reports the comparative results in terms of both average accuracy (AVG) across all tasks and final-task accuracy (LAST). Our method consistently outperforms all baselines. Specifically, compared to the top-performing methods in each scenario, our method (ViT-based) improves average accuracy improvements of 2.35%, 0.70%, and 0.96% on CIFAR-100 dataset, 1.03%, 0.54%, and 0.40% on TinyImageNet dataset, and 0.60% on ImageNet-Subset dataset. The ResNet-18 implementation also exhibits competitive performance, highlighting the effectiveness of our approach. Compared to the ResNet-18 variant, the ViT-based model shows superior results, suggesting the transformer architecture is well-suited for further exploration. Both variants maintain high final-task accuracy, mitigating catastrophic forgetting. Moreover, our method surpasses rehearsal-based methods using 20 exemplars per class, reinforcing the strength of our non-exemplar design.

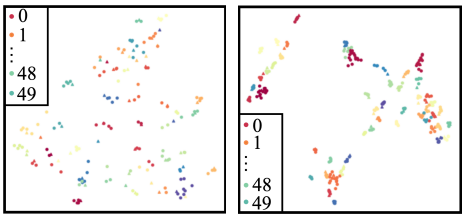
Accuracy Curves: Fig. 4 evaluates how accuracy varies as new tasks are added across different methods. While some methods outperform others in Task 0 accuracy, they exhibits accuracy drops after new tasks' done. In contrast, our method is more stable in accuracy, proving its superiority in balancing old and new tasks during incremental learning.

Forgetting Analysis: Tab. 2 compares average forgetting (AF) between our method with other approaches. As shown, our method has the lowest AF across all tasks, underscoring its strength in retaining historical knowledge and suitability for long-term tasks.

Confusion Matrix Analysis: We use the confusion matrix to analyze our DiAPR's performance after training on all tasks. As shown in Fig. 5, the matrix has few off-diagonal elements on historical classes, indicating strong old knowledge preservation. New classes show clear diagonal ele-

Method	Components	CIFAR-100		
		5 tasks	10 tasks	20 tasks
Baseline †	+ GaussAug	60.78	57.89	54.66
	+ NewFeatInj	62.13	60.78	57.02
	+DAP&GSA	63.47	61.67	57.98
Baseline † + **CDT**	+ GaussAug	62.82	60.12	57.69
	+ NewFeatInj	63.87	62.01	58.13
	+DAP&GSA	64.58	62.17	58.64

Table 3: Ablation study of DiAPR components on CIFAR-100 (LAST, %). GaussAug is the Gaussian augmentation and NewFeatInj means the New feature Injection.



(a) NewFeatInj (b) DiAPR

Figure 6: Visualization of historical prototypes after augmentation by New feature injection and DiAPR.

ments, demonstrating effective new knowledge learning.

Ablation Studies and Analysis

Component Effectiveness: We conduct ablation experiments on CIFAR-100 to evaluate the effectiveness of each component in DiAPR, with results presented in Tab. 3. Since DAP serves as a prerequisite for GSA, we integrated them into a combined module as DAP&GSA. Notably, the baseline that incorporates label augmentation with a ViT backbone, follows (Shi and Ye 2023; Li, Peng, and Zhou 2024; Zhai et al. 2024), has already demonstrated superior performance and supports its validity as a general-purpose baseline for further exploration. Our observations are as follows: (1) Compared to global-level augmentations (Gaussian noise and new feature injection), DAP&GSA performs dimension-wise semantic conflation. This enables prototypes to balance old class discriminability with adaptation to the new feature space. Experimental results confirm the higher performance of DAP&GSA. (2) CDT's dual functions not only additionally enhance prototypes' robustness but also improve feature space compatibility. Thus, CDT yields subsequent gains even with diverse prototype augmentation methods. (3) The DiAPR, integrating DAP, GSA, and CDT, delivers further performance gains. Through dimension-wise prototype refinement and softened feature alignment, DiAPR meets the stability-plasticity trade-off required in the NECIL setting.

Visualization of DiAPR: We perform t-SNE visualization (Van der Maaten and Hinton 2008) of the prototype distributions augmented via DiAPR and new feature injection to verify their discriminability when adapting to new feature spaces. As shown in Fig. 6, DiAPR yields more compact intra-class clusters, whereas global-level augmentation (NewFeatInj) induces dispersed distributions. These results

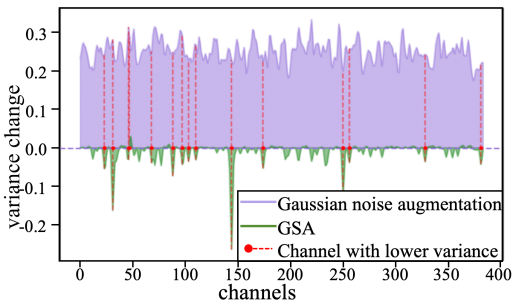


Figure 7: Dimension variance changes between original prototypes and those augmented by Gaussian noise or GSA.

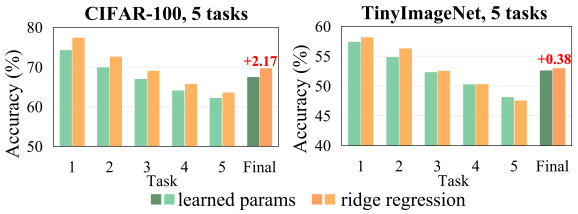


Figure 8: Accuracy comparison between ridge regression and learnable parameters across tasks.

confirm that our method effectively maintains old class discriminability in the evolving feature space.

Analysis of GSA: We further demonstrate our dimension-wise refinement by comparing dimensional variance changes between original prototypes and those augmented via our method versus Gaussian noise. As shown in Fig. 7, unlike Gaussian noise which uniformly perturbs all dimensions with roughly consistent variance changes, GSA allocates lower-variance dimensions (red marks) for new feature injection, gaining adaptability to the new feature space.

Analysis of CDT: We compare the effectiveness of ridge regression in CDT versus learnable parameters (fully connected layers) for constructing task transitions. As shown in Fig. 8, ridge regression outperforms learnable parameters on CIFAR-100 (+2.17%) and TinyImageNet (+0.38%). This likely benefits from: (1) Unlike learnable parameters, which are prone to sensitivity to random initialization and gradient fluctuations, ridge regression provides a stable closed-form solution across transitions. (2) Ridge regression explicitly embeds cross-dimensional dependencies, which enhances the prototype's robustness when adapting to new tasks.

Conclusion

We propose a DiAPR framework for NECIL. Specifically, historical prototypes are refined through three modules to sharpen classifier boundaries. DAP constructs semantically consistent pairs, which are fed into GSA for dimension-wise conflation, and CDT then enhances cross-dimensional dependencies. Additionally, CDT naturally yields softened feature alignment, resulting in a more compatible feature extractor. Extensive experiments validate the effectiveness of our method on different datasets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NO. 62176170, 62176169), Stability Program of National Key Laboratory of Security Communication (2024,WD202408), and the Sichuan Science and Technology Program (2025ZNSFSC0469).

References

Castro, F. M.; Marín-Jiménez, M. J.; Guil, N.; Schmid, C.; and Alahari, K. 2018. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 233–248.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, 532–547.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. Ieee.

Dong, S.; Gao, X.; He, Y.; Zhou, Z.; Kot, A. C.; and Gong, Y. 2024. CEAT: Continual Expansion and Absorption Transformer for Non-Exemplar Class-Incremental Learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Douillard, A.; Ramé, A.; Couairon, G.; and Cord, M. 2022. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9285–9295.

Gao, X.; He, Y.; Dong, S.; Cheng, J.; Wei, X.; and Gong, Y. 2023. Dkt: Diverse knowledge transfer transformer for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24236–24245.

Gomez-Villa, A.; Goswami, D.; Wang, K.; Bagdanov, A. D.; Twardowski, B.; and van de Weijer, J. 2024. Exemplar-free continual representation learning via learnable drift compensation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 473–490. Springer.

Goswami, D.; Liu, Y.; Twardowski, B.; and Van De Weijer, J. 2023. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems (NIPS)*, 36: 6582–6595.

Gu, Q.; Shim, D.; and Shkurti, F. 2023. Preserving linear separability in continual learning by backward feature projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24286–24295.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 831–839.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.(2009).

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, Q.; Peng, Y.; and Zhou, J. 2024. Fcs: Feature calibration and separation for non-exemplar class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28495–28504.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 2935–2947.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755. Springer.

Liu, X.; Zhai, J.-T.; Bagdanov, A. D.; Li, K.; and Cheng, M.-M. 2024. Task-adaptive saliency guidance for exemplar-free class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23954–23963.

Liu, Y.; Parisot, S.; Slabaugh, G.; Jia, X.; Leonardis, A.; and Tuytelaars, T. 2020. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, 699–716. Springer.

Liu, Y.; Schiele, B.; and Sun, Q. 2021. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2544–2553.

Malepathirana, T.; Senanayake, D.; and Halgamuge, S. 2023. Napa-vq: Neighborhood-aware prototype augmentation with vector quantization for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11674–11684.

Pelosin, F.; Jha, S.; Torsello, A.; Raducanu, B.; and van de Weijer, J. 2022. Towards exemplar-free continual learning in vision transformers: an account of attention, functional and weight regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3820–3829.

Petit, G.; Popescu, A.; Schindler, H.; Picard, D.; and Delezoide, B. 2023. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision (WACV)*, 3911–3920.

Qiu, Z.; Xu, Y.; Meng, F.; Li, H.; Xu, L.; and Wu, Q. 2024. Dual-Consistency Model Inversion for Non-Exemplar Class

Incremental Learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 24025–24035. IEEE Computer Society.

Rajasegaran, J.; Khan, S.; Hayat, M.; Khan, F. S.; and Shah, M. 2020. itaml: An incremental task-agnostic meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13588–13597.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001–2010.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*.

Roy, A.; Verma, V. K.; Voonna, S.; Ghosh, K.; Ghosh, S.; and Das, A. 2023. Exemplar-free continual transformer with convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5897–5907.

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

Schwarz, J.; Czarnecki, W.; Luketina, J.; Grabska-Barwinska, A.; Teh, Y. W.; Pascanu, R.; and Hadsell, R. 2018. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning (ICML)*, 4528–4537. PMLR.

Shi, W.; and Ye, M. 2023. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1772–1781.

Shi, Y.; Shi, D.; Qiao, Z.; Wang, Z.; Zhang, Y.; Yang, S.; and Qiu, C. 2023. Multi-granularity knowledge distillation and prototype consistency regularization for class-incremental learning. *Neural Networks*, 164: 617–630.

Sun, W.; Li, Q.; Zhang, J.; Wang, D.; Wang, W.; and Geng, Y.-a. 2023. Exemplar-free class incremental learning via discriminative and comparable parallel one-class classifiers. *Pattern Recognition*, 140: 109561.

Tang, S.; Chen, D.; Zhu, J.; Yu, S.; and Ouyang, W. 2021. Layerwise optimization by gradient decomposition for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 9634–9643.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, F.-Y.; Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2022. Foster: Feature boosting and compression for class-incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 398–414. Springer.

Yan, S.; Xie, J.; and He, X. 2021. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3014–3023.

Yu, L.; Twardowski, B.; Liu, X.; Herranz, L.; Wang, K.; Cheng, Y.; Jui, S.; and van de Weijer, J. 2020. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6980–6989.

Zhai, J.-T.; Liu, X.; Yu, L.; and Cheng, M.-M. 2024. Fine-grained knowledge selection and restoration for non-exemplar class incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 6971–6978.

Zhou, D.-W.; Wang, Q.-W.; Ye, H.-J.; and Zhan, D.-C. 2022. A model or 603 exemplars: Towards memory-efficient class-incremental learning. *arXiv preprint arXiv:2205.13218*.

Zhu, F.; Cheng, Z.; Zhang, X.-y.; and Liu, C.-l. 2021a. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems (NIPS)*, 34: 14306–14318.

Zhu, F.; Zhang, X.-Y.; Wang, C.; Yin, F.; and Liu, C.-L. 2021b. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 5871–5880.

Zhu, K.; Zhai, W.; Cao, Y.; Luo, J.; and Zha, Z.-J. 2022. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9296–9305.

Zhuang, H.; He, R.; Tong, K.; Zeng, Z.; Chen, C.; and Lin, Z. 2024. Ds-al: A dual-stream analytic learning for exemplar-free class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 17237–17244.