On the Discrimination and Consistency for Exemplar-Free Class Incremental Learning

Tianqi Wang[1,2], **Jingcai Guo**[1*], **Depeng Li**[3] and **Zhi Chen**[4]
[1]Department of COMP/LSGI, The Hong Kong Polytechnic University, Hong Kong SAR
[2]Department of Computer Science, University College London, United Kingdom
[3]School of AI and Automation, Huazhong University of Science and Technology, China
[4]School of Mathematics, Physics and Computing, The University of Southern Queensland, Australia
jc-jingcai.guo@polyu.edu.hk

Abstract

Exemplar-free class incremental learning (EF-CIL) is a nontrivial task that requires continuously enriching model capability with new classes while maintaining previously learned knowledge without storing and replaying any old class exemplars. An emerging theory-guided framework for CIL trains task-specific models for a shared network, shifting the pressure of forgetting to task-id prediction. In EF-CIL, task-id prediction is more challenging due to the lack of inter-task interaction (e.g., replays of exemplars). To address this issue, we conduct a theoretical analysis of the importance and feasibility of preserving a discriminative and consistent feature space, upon which we propose a novel method termed DCNet. Concretely, it progressively maps class representations into a hyperspherical space, in which different classes are orthogonally distributed to achieve ample inter-class separation. Meanwhile, it also introduces compensatory training to adaptively adjust supervision intensity, thereby aligning the degree of intra-class aggregation. Extensive experiments and theoretical analysis verified the superiority of DCNet. Code is available at https://github.com/Tianqi-Wang1/DCNet.

1 Introduction

Deep neural networks have achieved state-of-the-art performance in various tasks, yet they often struggle with Class Incremental Learning (CIL). In CIL, the model is constrained to learn new classes on non-stationary data distributions. This scenario can result in Catastrophic Forgetting (CF) [McCloskey and Cohen, 1989], as new parameters overwrite those learned for previous tasks. Meanwhile, CIL focuses on not relying on privileged information such as task-ids during inference [Wang *et al.*, 2023].

Exemplar-based methods, which preserve a portion of the samples from previous tasks for replay, have demonstrated strong performance in CIL. However, in the era of connectivity, data privacy has become increasingly crucial. The rising concern over data privacy conflicts with the exemplar-based approach [Zhuang *et al.*, 2022], thereby constraining its applicability. Recently, EF-CIL has attracted considerable attention because it entirely eliminates the need for replay samples, making it suitable for deployment in scenarios where privacy preservation and storage limitation are critical. Despite this advantage, existing EF-CIL methods are prone to more severe CF because training on a new task overwrites the parameter space learned for previous tasks. Classic strategies reduce the alteration of important weight by imposing constraints such as regularization using the Fisher information matrix [Kirkpatrick *et al.*, 2017]. More recent method leverages the Hilbert-Schmidt independence criterion for more stringent constraints [Li *et al.*, 2024b]. Rather than directly focusing on the weight, alternative approaches aim to maintain the semantic consistency of the prior feature space, typically through the use of class prototypes [Zhu *et al.*, 2021; Toldo and Ozay, 2022; Magistri *et al.*, 2024].

Parallel to these approaches, theoretical research suggests that a proficient CIL model can be broken down into a task-incremental learning (TIL) + out-of-distribution (OOD) task [Kim *et al.*, 2022c]. TIL typically involves training a separate model for each individual task and selecting appropriate inference paths or output heads through known task-ids. Based on this, the TIL+OOD architecture entails training a TIL-like model that constructs a new OOD classifier when faced with a new task. Consequently, an independent OOD classifier for each task would concurrently handle in-distribution (IND) classification, which is *within-task prediction*, and OOD detection to ensure accurate *task-id prediction*. During inference, for each test sample, the framework evaluates the probabilities of both to make a decision. This architecture facilitates the sharing of inter-task generalized knowledge while preserving intra-task specific knowledge, thereby demonstrating superior performance on CIL tasks. However, the TIL+OOD architecture results in task isolation during training, as it prevents access to the embedded representations and output magnitudes of previous tasks while learning the current task. Paradoxically, effective decision-making requires comparing outputs across the incremental sequence. This contradiction ultimately leads to a performance bottleneck in task-id prediction that requires inter-task interaction. Previous researches [Kim *et al.*, 2022b; Kim *et al.*, 2023; Lin *et al.*, 2024] have primarily focused on using replay samples to facilitate interaction, but this invades data privacy. Our

*Corresponding author: *Jingcai Guo*.

work emphasizes the implementation of inter-task interactions for the TIL+OOD framework in the context of EF-CIL.

From this perspective, we argue that preserving the discriminative and consistent feature space is crucial for enabling effective task interaction. *(i) Enhancing discriminability for a single task*. The feature space generated by general embedding methods, although effective for within-task prediction, frequently falls short in supporting OOD detection [Deng and Xiang, 2024; Ming *et al.*, 2023]. This limitation arises because OOD detection necessitates more discriminative features. *(ii) Ensuring consistency across incremental tasks*. Even with perfect OOD detection for each task, the isolation between tasks can lead to varying output magnitudes [Kim *et al.*, 2022b; Kim *et al.*, 2022c]. Our theoretical analysis indicates that maintaining discriminative and consistent feature space can be achieved by enhancing inter-class separation and aligning intra-class aggregation.

In this paper, we introduce a novel method to EF-CIL named Discriminative and Consistent Network (DCNet). This multi-head model leverages HAT [Serrà *et al.*, 2018] to learn task-specific masks for protecting the knowledge and makes decisions by comparing a sequence of OOD classifier outputs. To fully exploit the information in incremental tasks for interaction, DCNet comprises two key components: (1) Incremental Orthogonal Embedding (IOE), where we sequentially generate basis vectors that are orthogonally distributed on the unit hypersphere. The model then embeds the corresponding category features as closely as possible to these predefined vectors. This guarantees that the embedding of each category remains orthogonal to those of prior and future categories, thereby enhancing and aligning intra-task separation. (2) Dynamic Aggregation Compensation (DAC), which addresses the issue due to decreasing model plasticity by adaptively compensating for the reduced feature aggregation of subsequent tasks. DAC brings incremental feature embedding with more even intra-class aggregation. Benefiting from the synergy of these two components, DCNet effectively preserves the discriminative and consistent characteristics of the features, and does not rely on replaying samples or pre-trained models. Our main contributions are threefold:

- Theoretical analyses are provided to demonstrate the feasibility of leveraging information in incremental sequences to preserve discriminative and consistent features for the framework of TIL+ODD. To the best of our knowledge, we are the first to formally discuss how to optimize TIL+OOD model in the context of EF-CIL.
- DCNet not only incrementally embeds category features on the unit hypersphere but also maintains inter-class orthogonality. Furthermore, additional adaptive compensation helps balance the degree of intra-class aggregation across all tasks.
- Experiments conducted across multiple benchmark datasets consistently demonstrate that our method achieves highly competitive EF-CIL performance, with an average improvement of 8.33% over the latest state-of-the-art method on ImageNet-Subset task.

2 Related Work

Class-Incremental Learning. CIL necessitates that the model incrementally learn new classes without forgetting previously acquired knowledge. Classical CIL methods often maintain a certain number of exemplars from previous classes, which are replayed upon the arrival of a new task [Rebuffi *et al.*, 2017; Buzzega *et al.*, 2020; Yan *et al.*, 2021; Wang *et al.*, 2022; Wang *et al.*, 2023]. Replay strategy effectively reduces CF; however, concerns over privacy and memory limitation restrict its practicality. Exemplar-free approaches have focused on mitigating CF without relying on replay samples [Zhu *et al.*, 2023; Rypeść *et al.*, 2023; Gomez-Villa *et al.*, 2025]. EWC [Kirkpatrick *et al.*, 2017] employs the Fisher information matrix to constrain significant alterations in the weight space. LwF [Li and Hoiem, 2017] ensures that the output of the current model remains close to that of the previous model. PASS [Zhu *et al.*, 2021] leverages self-supervised learning to train a backbone network and maintains the consistency of class prototypes. FeTrIL [Petit *et al.*, 2023] transforms old prototype features based on the differences between old and new prototypes. ADC [Goswami *et al.*, 2024] uses adversarial samples against old task categories to estimate feature drift. EFC [Magistri *et al.*, 2024] identifies critical directions in the feature space for the previous task. However, some EF-CIL methods depend on a large initial task to train the backbone network and subsequently freeze it during increments. More recently, some studies have also explored utilizing pre-trained diffusion model or saliency detection network to mitigate forgetting [Meng *et al.*, 2025; Liu *et al.*, 2024]. Our end-to-end approach explores the application of TIL+OOD framework to EF-CIL without depending on a large initial task or a pre-trained model.

Task-id Predictor. One special approach to addressing the CIL problem involves utilizing multi-head models with task-id prediction. Specifically, CCG [Abati *et al.*, 2020] builds a separate network to predict task-id, while iTAML [Rajasegaran *et al.*, 2020] necessitates batched samples for task-id prediction during inference. HyperNet [Von Oswald *et al.*, 2019] and PR-Ent [Henning *et al.*, 2021] employ entropy for task-id prediction. Prior research has highlighted that the performance bottleneck of these systems stems from failing to realize the relationship between task-id prediction and OOD detection [Kim *et al.*, 2022c]. OOD detection requires models not only to accurately identify data from known distributions (i.e., categories learned during training), but also to detect samples outside these distributions (i.e., unknown categories) [Morteza and Li, 2022; Ming *et al.*, 2023; Lu *et al.*, 2024]. Kim et al. [2022c] conducted a theoretical analysis of the TIL+OOD architecture, demonstrating its applicability to CIL tasks. Building on this, MORE [Kim *et al.*, 2022b] and ROW [Kim *et al.*, 2023] adopt the same structure using pre-trained models and replay samples. The latest work TPL [Lin *et al.*, 2024], leverages replay samples to construct likelihood ratios, thereby enhancing task-id prediction. Our proposed method also falls under the TIL+OOD framework. However, unlike previous methods, DCNet focuses on efficiently utilizing information from incremental sequences to accomplish inter-task interaction without replay samples.

3 Theoretical Analysis

3.1 Class-Incremental Learning Setup

Class-Incremental Learning (CIL) aims to address a sequence of tasks $1, \ldots, T$. Each task t consists of an input space $\mathcal{X}^{(t)}$, a label space $\mathcal{Y}^{(t)}$, and a training set $\mathcal{D}^{(t)} = \{(x_j^{(t)}, y_j^{(t)})\}_{j=1}^{N^{(t)}}$, where $N^{(t)}$ is the number of samples. The label spaces of different tasks have no overlap, i.e., $\mathcal{Y}^{(i)} \cap \mathcal{Y}^{(k)} = \emptyset, \forall i \neq k$. The objective of CIL is to train a progressively updated model that can effectively map the entire input space $\bigcup_{t=1}^{T} \mathcal{X}^{(t)}$ to the corresponding label space $\bigcup_{t=1}^{T} \mathcal{Y}^{(t)}$. Kim et al. [2022c] proposed a novel theory for solving the CIL problem. They decomposed the probability of a sample x belonging to class $y_j^{(t)}$ of task t as follows:

$$P(y_j^{(t)} \mid x) = P(y_j^{(t)} \mid x, t) \cdot P(t \mid x). \tag{1}$$

The formulation can be decoupled into two components: *within-task prediction* and *task-id prediction*. In CIL, data from different tasks can be considered as OOD samples to each other. However, only relying on traditional OOD detection methods to build separate model results in task isolation. We highlight that interaction between tasks can be achieved by preserving the discriminative and consistent feature space. In the subsequent sections, we analyze how these properties can be implemented theoretically. Subsection 3.2 underscores the importance of inter-class separation and intra-class aggregation through a theorem. Subsection 3.3 elaborates on how both concepts facilitate information interaction.

3.2 OOD Detection Capabilities

We now discuss how inter-class separation and intra-class aggregation affect the performance of OOD detection in the context of EF-CIL. The theory proposed by Morteza and Li [2022] demonstrates that the effectiveness of pure OOD detection is closely tied to the distance between the IND and OOD data. We generalize this theory to incremental learning tasks.

Let $\{\mu_{\text{in},i}\}_{i=1}^{k}$ be the mean vectors of k Gaussian components representing the IND (a task contain k distinct categories). Consider a sequence of OOD Gaussian mean vectors $\{\mu_{\text{out},t}\}_{t=1}^{T}$, describing a uniformly weighted Gaussian mixture model for OOD data (an incremental sequence contain T tasks), and let Σ be the positive definite shared covariance matrix. IND and OOD exhibit distinct and significant differences. Define a scoring function $ES(x)$ that is proportional to the data density, and a measure D of the expectation difference in $ES(x)$ between IND and OOD samples:

$$ES(x) \;=\; \sum_{i=1}^{k} \exp\left(-\tfrac{1}{2}\,(x-\mu_i)^\top \Sigma^{-1}\,(x-\mu_i)\right), \tag{2}$$

$$D \;=\; \mathbb{E}_{x\sim P_{\mathcal{X}}^{\text{in}}}\big[ES(x)\big] \;-\; \mathbb{E}_{x\sim P_{\mathcal{X}}^{\text{out}}}\big[ES(x)\big]. \tag{3}$$

Recall the Mahalanobis distance: $d_M(u,v) = \sqrt{(u-v)^\top \Sigma^{-1}\,(u-v)}$. Our objective is to investigate the factors influencing D, a metric that quantifies the performance of OOD detection in incremental sequences. We investigate the upper bound of D, which is derived using the Total Variation and the Pinsker's inequality (see Lemma 3 and 5).

Lemma 1. *Let* $\alpha_{i,t} := \frac{1}{2}\, d_M\big(\mu_{\text{in},i}, \mu_{\text{out},t}\big)$, $i = 1, \ldots, k,\ t = 1, \ldots, T$, *then we have the following estimate:*

$$\mathbb{E}_{x\sim P_{\mathcal{X}}^{in}}(ES(x)) - \mathbb{E}_{x\sim P_{\mathcal{X}}^{out}}(ES(x)) \leq \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{k}\alpha_{i,t}. \tag{4}$$

Details of the proof are provided in Appendix A.

Lemma 1 demonstrates that as $\mu_{\text{in},i}$ and $\mu_{\text{out},t}$ become more distant, the overall OOD detection performance improves. However, within the context of EF-CIL, the OOD data is unavailable, making it challenging to explicitly increase the divergence. Consequently, the subsequent derivation aims to further relax the upper bound while introducing inter-class separation, which can be estimated in EF-CIL.

For any $\mu_{\text{out},t}$, we can always find a nearest μ_{in,i_0} that satisfies the triangle inequality, denoted as:

$$d_M(\mu_{\text{in},i}, \mu_{\text{out},t}) \leq d_M(\mu_{\text{out},t}, \mu_{\text{in},i_0}) + d_M(\mu_{\text{in},i_0}, \mu_{\text{in},i}). \tag{5}$$

We end up with the following theorem.

Theorem 1. *Consider a sequential OOD detection task in the context of CIL, we have the following bounds:*

$$\begin{aligned}&\mathbb{E}_{x\sim P_{\mathcal{X}}^{\text{in}}}\big[ES(x)\big] - \mathbb{E}_{x\sim P_{\mathcal{X}}^{\text{out}}}\big[ES(x)\big] \\ &\leq \frac{k}{2T}\sum_{t=1}^{T} d_M\big(\mu_{\text{out},t}, \mu_{\text{in},i_0}\big) + \frac{1}{2}\sum_{i=1}^{k} d_M\big(\mu_{\text{in},i_0}, \mu_{\text{in},i}\big).\end{aligned} \tag{6}$$

Theorem 1 handles incremental task arrivals and introduces inter-class separation $d_M(\mu_{\text{in},i_0}, \mu_{\text{in},i})$. It can be found that the performance of TIL+OOD approaches can be improved by increasing the inter-class difference of IND prototypes. In essence, if the IND data are well-separated, more space exists in the feature space for OOD samples to be embedded.

Finally, we examine how intra-class aggregation influences detection performance through the shared covariance matrix Σ. If Σ becomes "smaller" in the positive-definite ordering (i.e., $\Sigma_a \preceq \Sigma_b$ implies $\Sigma_b - \Sigma_a$ is positive semidefinite), then Σ_a^{-1} is "larger" compared to Σ_b^{-1}. Consequently, the same Euclidean displacement leads to a larger Mahalanobis distance under a smaller covariance. For any fixed vector $(u-v)$, if $\Sigma_a \preceq \Sigma_b$, then:

$$(u-v)^\top \Sigma_a^{-1}\,(u-v) \;\;\geq\;\; (u-v)^\top \Sigma_b^{-1}\,(u-v). \tag{7}$$

Therefore, when the IND data exhibits a higher degree of intra-class aggregation, OOD samples are pushed farther away in the Mahalanobis distance sense.

Interpretation. Our analysis highlights two crucial factors. **Inter-class separation:** A larger separation between IND prototypes offers a better theoretical margin, especially when the OOD dynamics changes and cannot be estimated. **Intra-class aggregation:** A smaller covariance matrix Σ indicates that the IND data are more tightly clustered around the class prototype, thereby amplifying Mahalanobis distances to potential OOD samples.

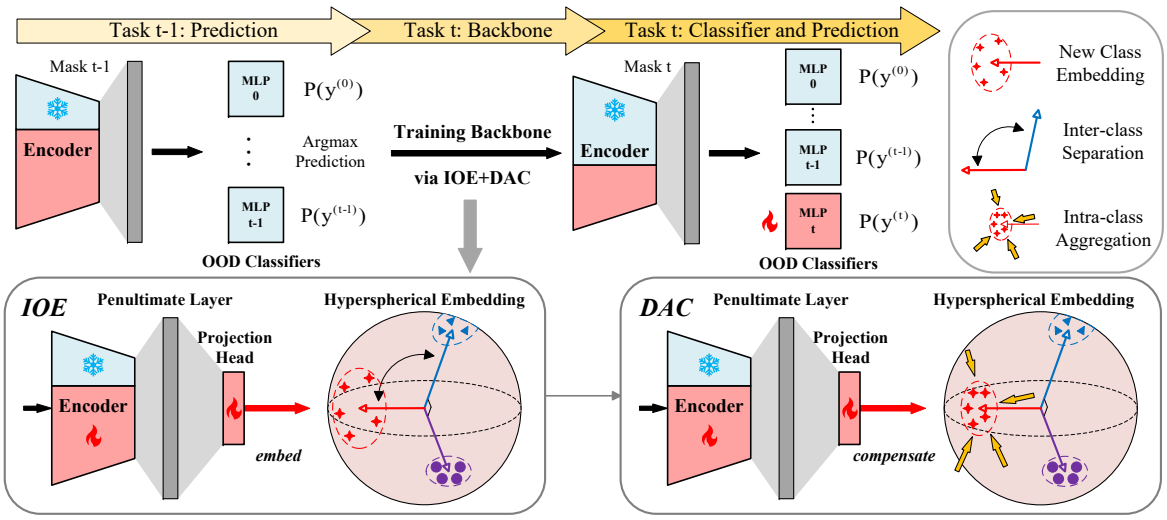

Figure 1: Overview of DCNet. Upon the arrival of task t, DCNet optimizes the learnable part of the backbone and creates a new OOD classifier. Through IOE, DCNet incrementally embeds new features in directions that remain orthogonal to previous categories. Subsequently, the DAC module dynamically compensates for any insufficient aggregation by referencing the degree of aggregation from the previous tasks.

3.3 Task Information Interaction

In this subsection, we will discuss the feasibility of employing inter-class separation and intra-class aggregation to preserve discriminative and consistent feature spaces. The TIL+OOD approach consists of a backbone network trained using TIL-like methods and multiple OOD classifiers. During the inference, the final decision $\hat{y}$ is determined by selecting the highest output among the T OOD classifiers:

$$\hat{y} = \arg \max_{1 \le t \le T} \oplus P(y^{(t)} \mid x, t), \tag{8}$$

where $\oplus$ denotes the concatenation over the output space. We select the class with the highest softmax probability over each task among all the learned classes. However, even if each individual OOD decision is perfect, Eq. (8) may still lead to incorrect CIL predictions due to varying magnitudes of outputs across different tasks. Breaking the isolation of tasks during the training is essential for achieving comparable outputs. Previous studies [Kim *et al.*, 2022b; Kim *et al.*, 2023; Lin *et al.*, 2024] have facilitated direct information interaction via replay samples; however, these approaches are constrained in the context of EF-CIL. As analyzed in Subsection 3.2, inter-class separation and intra-class aggregation are critical for maintaining the validity and integrity of the feature space without violating the privacy. Specifically, inter-class separation preserves the discrimination of the feature space, while intra-class aggregation ensures the consistency, thereby enabling effective task information interaction.

4 Methodology

4.1 Overview

DCNet comprises two essential components: Incremental Orthogonal Embedding (IOE) and Dynamic Aggregation Compensation (DAC), which operate synergistically (see Figure 1). Consistent with previous works, our approach employs the mask-based method HAT [Serrà *et al.*, 2018] to shift CF. Specifically, as each task is learned, the model generates a set of masks for important neurons, ensuring that these masks are as compact as possible. Formally, we introduce a loss term $\mathcal{L}_{HAT}$. During the learning of a new task, the masks from the previous model inhibit backpropagation from updating the masked neurons. Since all neurons remain accessible during the forward propagation, inter-task generalized knowledge can be leveraged across all tasks. As the mask is progressively learned, the available weights for updating by subsequent tasks become increasingly sparse, leading to a decrease in model plasticity. This results in the degradation of the feature space for subsequent tasks, where later tasks often exhibit poorer performance compared to earlier tasks in terms of inter-class separation and intra-class aggregation. Through IOE and DAC components, DCNet preserves a discriminative and consistent feature space by facilitating inter-task information interaction.

4.2 Incremental Orthogonal Embedding (IOE)

The core principle of IOE lies in explicitly associating each category with incrementally generated basis vectors that maintain orthogonality, thereby ensuring superior inter-class separation. For an incremental task t with data $\mathbf{x}^{(t)}$, the framework consists of two mappings: An encoder $f : \mathcal{X} \to \mathbb{R}^f$, which maps the input $\mathbf{x}^{(t)}$ to a feature $\mathbf{f}^{(t)} = f(\mathbf{x}^{(t)})$. A projection head $h : \mathbb{R}^f \to \mathbb{R}^z$, which further maps $\mathbf{f}^{(t)}$ to a embedding $\tilde{\mathbf{z}}^{(t)} = h(\mathbf{f}^{(t)})$. The output embeddings are normalized as $\mathbf{z}^{(t)} = \tilde{\mathbf{z}}^{(t)} / \|\tilde{\mathbf{z}}^{(t)}\|^2$ to reside on a unit hypersphere.

We aim to more uniformly distribute the unit hypersphere and constrain individual category features within their respective spaces to minimize overlap between categories [Deng and Xiang, 2024]. The approach to updating class prototypes is data-driven, which cannot control the location of class prototypes [Ming *et al.*, 2023]. Based on this, we argue that binding category features to corresponding basis vectors to enforce orthogonality is essential for maintaining dis-

tinct boundaries and minimizing interference between different classes. By ensuring 90° angular separation between basis vectors, each category can occupy a unique region in the feature space. Furthermore, this component explicitly defines the placement of new class embeddings, thereby preventing potential reductions in inter-class separation that could result from diminished model plasticity.

We design an incremental generator that produces predefined, mutually orthogonal basis vectors through a data-independent process. Let $\{\mu_c^{\text{old}}\}_{c=1}^{C^{\text{old}}}$ be the existing set of C^{old} basis vectors, each normalized to unit length and mutually orthogonal. When new classes $C^{(t)}$ of task t arrive, each new generated vector $\{\mu_k^{(t)}\}_{k=1}^{C^{(t)}}$ is also normalized and required to be orthogonal to both existing and newly added vectors. To approximately satisfy these constraints, let $M^* = \{\mu_c^{\text{old}}, \mu_k^{(t)}\}$ represent all basis vectors after incrementally adding new classes, and define the following objective:

$$M^* = \arg\min_M \left[\sum_{i,j=1}^{C^{(t)}} |{\mu_i^{(t)}}^\top \mu_j^{(t)}| + \sum_{k=1}^{C^{(t)}} \sum_{c=1}^{C^{\text{old}}} |{\mu_k^{(t)}}^\top \mu_c^{\text{old}}| \right], \tag{9}$$

where each pairwise inner product deviating from zero is penalized. During the generation of new basis vectors, the previous basis vectors are kept unchanged. Subsequently, by employing Eq. (9) as the optimization objective, the new basis vectors are updated through gradient descent to enforce mutual orthogonality.

Having obtained the necessary vectors, we should now concentrate on embedding features in proximity to these basis vectors. Since these basis vectors are distributed on the unit hypersphere, we can model the embedding effectively using the von Mises-Fisher (vMF) distribution [Mardia and Jupp, 2009]. The vMF distribution serves as a spherical counterpart to Gaussian distributions, designed for unit norm embeddings $\mathbf{z}$ where $\|\mathbf{z}\|^2 = 1$. The probability density function of a unit vector $\mathbf{z} \in \mathbb{R}^d$ belonging to class k in task t is defined as:

$$p_d(\mathbf{z}^{(t)}; \mu_k^{(t)}, \kappa) = Z_d(\kappa) \exp\left(\kappa \mathbf{z}^{(t)} \mu_k^{(t)}\right), \tag{10}$$

where $\mu_k^{(t)}$ is the generated basis vector for class k in M^*, $\kappa \geq 0$ represents the concentration parameter controlling the distribution tightness, and $Z_d(\kappa)$ is the normalization factor. The optimized normalized probability of assigning an embedding $\mathbf{z}_i^{(t)}$ to category $c_{(i)}$ is given as follows:

$$\mathcal{L}_{\text{IOE}} = -\frac{1}{N^{(t)}} \sum_{i=1}^{N^{(t)}} \log \frac{\exp\left(\mathbf{z}_i^{(t)} \mu_{c_{(i)}}^{(t)} / \tau_{\text{IOE}}\right)}{\sum_{j=1}^{C^{(t)}} \exp\left(\mathbf{z}_i^{(t)} \mu_j^{(t)} / \tau_{\text{IOE}}\right)}, \tag{11}$$

where $c_{(i)}$ denotes the class index of a sample x_i in task t, τ_{IOE} is the fixed temperature. Combining Eqs. (9) and (11), IOE successfully embeds categories orthogonally on the unit hypersphere. Following previous researches [Zhu *et al.*, 2021; Kim *et al.*, 2022a; Magistri *et al.*, 2024], we also employed self-rotation augmentation. It is crucial to highlight that IOE communicates an important information to each task: to precisely delineate inter-class separation and minimize overlap. This information interaction facilitates superior feature discrimination.

4.3 Dynamic Aggregation Compensation (DAC)

DAC focuses on the degree of intra-class aggregation. Specifically, as model plasticity decreases, the embeddings for subsequent tasks tend to become increasingly diffuse. To counteract this, DAC employs the aggregation patterns from previous tasks as a template, dynamically adjusting pressure to maintain consistent aggregation. We introduce and minimize an adaptive supervised contrastive loss to compensate for inadequate aggregation after the IOE has undergone a predefined number of training iterations:

$$\mathcal{L}_{\text{DAC}} = -\frac{1}{N^{(t)}} \sum_{i=1}^{N^{(t)}} \frac{1}{|P(i)^{(t)}|} \times \sum_{p \in P(i)^{(t)}} \log \frac{\exp(\mathbf{z}_i^{(t)} \cdot \mathbf{z}_p^{(t)} / \tau^{(t)})}{\sum_{j=1, j \neq i}^{N} \exp(\mathbf{z}_i^{(t)} \cdot \mathbf{z}_j^{(t)} / \tau^{(t)})}, \tag{12}$$

where $P(i)^{(t)}$ is the set of positive samples for sample x_i, $\mathbf{z}_i^{(t)}$ and $\mathbf{z}_p^{(t)}$ are the embedding representations of sample x_i and its positive sample x_p, $\tau^{(t)}$ is an adaptive temperature that controls the compensation intensity. Adjusting temperature to optimize training for a single task is common; however, DAC aims to leverage information from incremental tasks to select the optimal $\tau^{(t)}$ for precise compensation. Indeed, employing fixed hyperparameters in CIL is suboptimal, as appropriate adjustments are necessary based on factors such as task complexity [Semola *et al.*, 2024; Li *et al.*, 2024a].

Naturally, we can estimate the concentration parameter κ based on the distributional form of Eq. (10), and subsequently employ this estimate as a degree of aggregation. However, due to the presence of the Bessel function, an analytic solution for κ is not feasible. In DAC, we adopt a more intuitive approach by calculating the average cosine similarity between samples and basis vector to quantify the degree of aggregation $\omega^{(t)}$, corresponding to the exponential term in the vMF distribution. This measure is then used to dynamically adjust the temperature relative to the historical average degree of aggregation ω^{avg}:

$$\omega^{(t)} = \frac{1}{N^{(t)}} \sum_{i=1}^{N^{(t)}} \mathbf{z}_i^{(t)} \cdot \mu_{c_{(i)}}^{(t)}, \quad \tau^{(t)} = \tau^{(0)} \cdot \frac{\omega^{(t)}}{\omega^{\text{avg}}}. \tag{13}$$

DAC calculates the aggregation degree $\omega^{(t)}$ via Eq. (13) over a specified epoch and subsequently updates the temperature $\tau^{(t)}$ for Eq. (12). If the aggregation degree of the current task is insufficient, DAC applies a lower temperature to enhance the embedding intensity; conversely, if the aggregation degree is adequate, the constraint is relaxed appropriately. This balancing mechanism uses the degree of intra-class aggregation as information to flow between tasks, thereby ensuring that each class maintains a consistent and compact embedding throughout the incremental learning process.

In summary, IOE and DAC enhance the inter-class separation and align the intra-class aggregation, constructing a discriminative and consistent feature space through the interaction between tasks. The final optimization objective comprises three components:

Method	CIFAR-100				Tiny-ImageNet				ImageNet-Subset			
	Split-10		Split-20		Split-10		Split-20		Split-10		Split-20	
	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}
EWC	49.14	31.17	31.02	17.37	24.01	8.00	15.70	5.16	39.40	24.59	26.95	12.78
LwF	53.91	32.80	38.39	17.44	45.14	26.09	32.94	15.02	56.41	37.71	40.23	18.64
PASS	47.86	30.45	32.86	17.44	39.25	24.11	32.01	18.73	45.74	26.40	31.65	14.38
FeTrIL	51.20	34.94	38.48	23.28	45.60	30.97	39.54	25.70	52.63	36.17	42.43	26.63
SSRE	47.26	30.40	32.45	17.52	38.82	22.93	30.62	17.34	43.76	25.42	31.15	16.25
EFC	58.58	43.62	47.36	32.15	*47.95*	34.10	*42.07*	*28.69*	59.94	47.38	49.92	35.75
LDC	59.50	45.40	-	-	46.80	*34.20*	-	-	*69.40*	51.40	-	-
ADC	61.35	46.48	-	-	43.04	32.32	-	-	67.07	46.58	-	-
SEED	*62.04*	*51.42*	*57.42*	*42.87*	-	-	-	-	67.55	*55.17*	*62.26*	*45.77*
DCNet	**75.84** ±0.52	**65.40** ±0.26	**71.52** ±0.42	**58.43** ±0.36	**57.00** ±0.22	**48.37** ±0.33	**50.05** ±0.10	**36.75** ±0.29	**76.82** ±0.25	**67.82** ±0.22	**69.12** ±0.43	**50.31** ±0.53

Table 1: Comparison with baselines on Split CIFAR-100, Tiny-ImageNet, and ImageNet-Subset. All methods are trained from scratch without using replay samples. Our method is evaluated over five runs, with the mean performance and standard deviation reported. We emphasize the optimal results in bold and denote the sub-optimal results in italics.

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{IOE}} + \lambda \cdot \mathcal{L}_{\text{DAC}} + \lambda_{\text{HAT}} \cdot \mathcal{L}_{\text{HAT}}, \tag{14}$$

where λ and λ_{HAT} are hyper-parameters used to balance the total loss. The algorithm for HAT and the procedure for DCNet are provided in Appendix B. We also discuss the connection between the algorithm and the theoretical analysis.

5 Experiment

5.1 Experiment Setting

Datasets. For a fair comparison with baselines, we utilize three widely adopted datasets in CIL. The CIFAR-100 [Krizhevsky *et al.*, 2009] comprises 50k training images and 10k test images, each sized 32×32 pixels, spanning 100 categories. The Tiny-ImageNet [Le and Yang, 2015], a subset of ImageNet, includes 100k training images and 10k test images, each sized 64×64 pixels, covering 200 categories. The ImageNet-Subset is a subset of the ImageNet (ILSVRC 2012) [Russakovsky *et al.*, 2015] with 100 categories, containing approximately 130k training images, each sized 224×224 pixels. We split these datasets equally into 10-task and 20-task sequences. This experimental setup is more challenging and realistic because it does not rely on a large initial task.

Baselines. Since our focus is on the EF-CIL scenario, we conduct comprehensive comparisons with both classical and state-of-the-art EF-CIL methods: **EWC** [Kirkpatrick *et al.*, 2017], **LwF** [Li and Hoiem, 2017], **PASS** [Zhu *et al.*, 2021], **FeTrIL** [Petit *et al.*, 2023], **SSRE** [Zhu *et al.*, 2023], **EFC** [Magistri *et al.*, 2024], **LDC** [Gomez-Villa *et al.*, 2025], **ADC** [Goswami *et al.*, 2024], **SEED** [Rypeść *et al.*, 2023]. Furthermore, to substantiate the effectiveness of our approach, we also compare it with several exemplar-based methods, particularly various TIL+OOD approaches, including: **iCaRL** [Rebuffi *et al.*, 2017], **DER++** [Buzzega *et al.*, 2020], **DER** [Yan *et al.*, 2021], **FOSTER** [Wang *et al.*, 2022], **BEEF** [Wang *et al.*, 2023], **MORE** [Kim *et al.*, 2022b], **ROW** [Kim *et al.*, 2023], **TPL** [Lin *et al.*, 2024].

Training. We employ a ResNet-18 model [He *et al.*, 2016] trained from scratch for all experiments. For comparison baselines, we either reproduce the results using the hyper-parameters specified in their source code repositories or directly adopt the existing results in state-of-the-art baselines. To ensure a fair comparison, we also allow the baselines to utilize self-rotation augmentation [Magistri *et al.*, 2024]. For CIFAR-100 and Tiny-ImageNet, consistent with prior work [Kim *et al.*, 2022a], we utilize LARS [You *et al.*, 2017] training for 700 epochs with an initial learning rate of 0.1, introducing the DAC component at epoch 400. For ImageNet-Subset, we train for 100 epochs, incorporating DAC at epoch 50. In all the experiments, we set the $\tau_{\text{IOE}} = 0.05$, $\tau^{(0)} = 0.2$ and configure the dimension of basis vector be 256.

Evaluation. We report two key metrics: $A_{\text{last}}(\%)$, which represents the average accuracy after the last task; $A_{\text{inc}}(\%)$, which denotes the average incremental accuracy across all tasks. For further details on training and metric calculations, please refer to Appendix C.

5.2 Main Comparison Results

Comparison with EF-CIL Approaches. Table 1 offers a comprehensive comparison of various baselines across three standard benchmark datasets. All methods were trained from scratch without utilizing any replay samples. Our method demonstrates significant performance among all baselines with non-marginal improvements. Specifically, compared to the second-best method, DCNet achieves improvements of 11.01% and 11.68% in average and final accuracy metrics, respectively. This underscores the superior competitiveness of DCNet within EF-CIL. This outstanding performance can be attributed to our improvement of the TIL+OOD framework, which fundamentally differs from the traditional EF-CIL approach and effectively overcomes the isolation between tasks.

Comparison with Exemplar-based Approaches. Table 2 presents a comparative analysis of our approach against several state-of-the-art exemplar-based methods. All baselines train from scratch while maintaining a buffer $\mathcal{M}$ of 2000 samples. DCNet achieves a performance improvement of 0.9% and 5.5% over these advanced methods on two 10 tasks sequences. In the ImageNet-Subset-split10 task, our method

Method	$\mathcal{M}$	CIFAR100	Tiny-ImageNet	ImageNet-Subset
iCaRL		51.4	28.3	50.98
DER++		53.7	30.5	-
DER		*64.5*	38.3	66.85
FOSTER		62.5	36.4	67.68
BEEF	2k	60.9	37.9	**68.78**
MORE†		57.5	35.4	-
ROW†		58.2	38.2	-
TPL†		62.2	*42.9*	-
DCNet	0	**65.4**	**48.4**	*67.82*

Table 2: Comparison with exemplar-based baselines, where we report the average accuracy denoted as A_{last}. †: These methods also adopt the same TIL+OOD framework.

Component	λ	CIFAR100-10		ImageNet100-10	
		A_{inc}	A_{last}	A_{inc}	A_{last}
HAT+CSI	-	73.30	63.32	70.80	63.94
IOE	-	73.85	63.80	73.55	64.86
IOE+DAC†	1.0	74.04	64.40	75.09	65.80
IOE+DAC††	0.5	75.49	65.27	76.19	66.68
IOE+DAC††	1.0	**75.84**	**65.40**	**76.82**	**67.82**
IOE+DAC††	2.0	75.58	64.92	75.64	65.92

Table 3: Effectiveness of the core designs in our DCNet. HAT+CSI serves as the foundation for our approach. †: DAC component with fixed temperature; ††: DAC component with dynamic temperature.

slightly underperforms the state-of-the-art baseline. It is important to note that maintaining a large buffer is an effective strategy for complex tasks. Specifically, buffering 297 samples from ImageNet consumes memory equivalent to that of a ResNet-18 backbone network. Therefore, a fair comparison should account for these resource requirements [Zhou *et al.*, 2023]. Furthermore, we highlight that MORE, ROW, and TPL are also TIL+OOD methods that explicitly use replay samples to break task isolation. In contrast, our approach constructs the interaction of information among tasks without replay samples, and achieves performance that is competitive with exemplar-based methods.

5.3 Algorithm Analysis

We evaluate the empirical effectiveness of DCNet. Table 3 presents the results of the ablation study and parametric analysis. HAT+CSI, a pioneering approach for TIL+OOD, is introduced by Kim et al. [2022c] to demonstrate the feasibility of the TIL+OOD framework; however, it does not consider the information interaction between tasks. Consequently, HAT+CSI can be considered a precursor to DCNet.

Building on this, Table 3 provides results for three configurations: IOE, using only the IOE component; IOE+DAC†, combining the fixed DAC component with IOE; IOE+DAC††, combining the dynamic DAC component with IOE using different λ for Eqs. (14). The results indicate that our methods are interdependent and achieve superior performance by fully utilizing available information. It is noteworthy that the performance gains of DCNet are particularly pronounced in the more complex ImageNet-Subset task.

Figure 2 visualizes the feature space of a single task, where (a) (b) present the results of HAT+CSI and (c) (d) represent our method. Each group visualizes three distinct classes in a three-dimensional space and an entire task in a two-dimensional space. Attribute to the superior inter-class separation by IOE, the feature space of DCNet is more discriminative. Figure 3 illustrates the decreasing trend of intra-class aggregation $\omega^{(t)}$ as tasks progress, due to diminished model plasticity. To counteract this effect, the DAC component dynamically adjusts the compensation intensity based on the degree of aggregation from previous tasks, thereby stabilizing the change curve of intra-class aggregation. Finally, Figure 4 presents the average inter-class Mahalanobis distance across the incremental sequence. By leveraging the two components, DCNet constructs a feature space that is both discriminative and consistent.

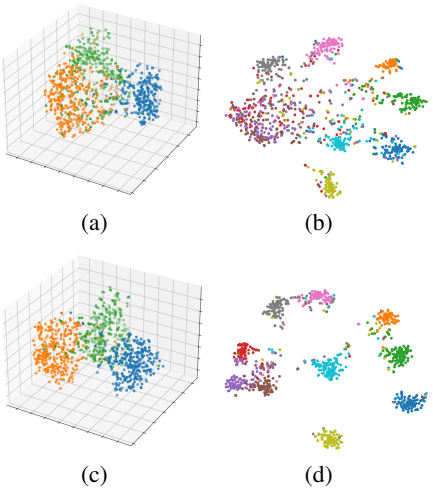

Figure 2: t-SNE visualization of the embedding space, where a color represents a category. (a) (b) illustrate the embedding results for HAT+CSI; and (c) (d) present results for our method.

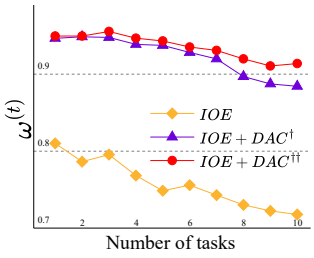

Figure 3: Changes in the degree of aggregation $\omega^{(t)}$.

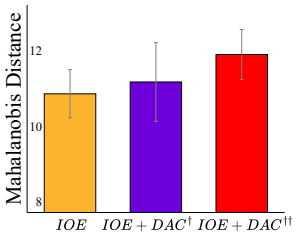

Figure 4: Average Mahalanobis distance between classes.

6 Conclusion

We introduce a novel approach of TIL+OOD framework in the context of EF-CIL, leveraging information from incremental sequence to overcome task isolation. Theoretical analysis reveals that inter-class separation and intra-class aggregation are crucial for effective OOD detection in an incremental sequence. Our proposed DCNet, informed by these insights, preserves the discriminability and consistency of the feature space via its IOE and DAC components. Extensive experiments validate the competitiveness of our method. Future research could explore scenarios with blurry task boundaries.

Acknowledgements

This research was supported by funding from the Hong Kong RGC General Research Fund (152211/23E, 15216424/24E, and 152115/25E), the National Natural Science Foundation of China (623B2040), the PolyU Internal Fund (P0056171) and the Huawei Gifted Fund.

References

[Abati *et al.*, 2020] Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3931–3940, 2020.

[Buzzega *et al.*, 2020] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*, 33:15920–15930, 2020.

[Deng and Xiang, 2024] Yao Deng and Xiang Xiang. Expanding hyperspherical space for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1967–1976, 2024.

[Gomez-Villa *et al.*, 2025] Alex Gomez-Villa, Dipam Goswami, Kai Wang, Andrew D Bagdanov, Bartlomiej Twardowski, and Joost van de Weijer. Exemplar-free continual representation learning via learnable drift compensation. In *European Conference on Computer Vision*, pages 473–490. Springer, 2025.

[Goswami *et al.*, 2024] Dipam Goswami, Albin Soutif-Cormerais, Yuyang Liu, Sandesh Kamath, Bart Twardowski, Joost van de Weijer, et al. Resurrecting old classes with new data for exemplar-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28525–28534, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Henning *et al.*, 2021] Christian Henning, Maria Cervera, Francesco D'Angelo, Johannes Von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F Grewe, and Joao Sacramento. Posterior meta-replay for continual learning. *Advances in Neural Information Processing Systems*, 34:14135–14149, 2021.

[Kim *et al.*, 2022a] Gyuhak Kim, Sepideh Esmaeilpour, Changnan Xiao, and Bing Liu. Continual learning based on ood detection and task masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3856–3866, 2022.

[Kim *et al.*, 2022b] Gyuhak Kim, Bing Liu, and Zixuan Ke. A multi-head model for continual learning via out-of-distribution replay. In *Conference on Lifelong Learning Agents*, pages 548–563. PMLR, 2022.

[Kim *et al.*, 2022c] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. *Advances in Neural Information Processing Systems*, 35:5065–5079, 2022.

[Kim *et al.*, 2023] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, and Bing Liu. Learnability and algorithm for continual learning. In *International Conference on Machine Learning*, pages 16877–16896. PMLR, 2023.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 2009.

[Le and Yang, 2015] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[Li and Hoiem, 2017] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

[Li *et al.*, 2024a] Depeng Li, Tianqi Wang, Junwei Chen, Wei Dai, and Zhigang Zeng. Harnessing neural unit dynamics for effective and scalable class-incremental learning. In *International Conference on Machine Learning*, pages 28688–28705, 2024.

[Li *et al.*, 2024b] Depeng Li, Tianqi Wang, Junwei Chen, Qining Ren, Kenji Kawaguchi, and Zhigang Zeng. Towards continual learning desiderata via hsic-bottleneck orthogonalization and equiangular embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13464–13473, 2024.

[Lin *et al.*, 2024] Haowei Lin, Yijia Shao, Weinan Qian, Ningxin Pan, Yiduo Guo, and Bing Liu. Class incremental learning via likelihood ratio based task prediction. In *International Conference on Learning Representations*, 2024.

[Liu *et al.*, 2024] Xialei Liu, Jiang-Tian Zhai, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Task-adaptive saliency guidance for exemplar-free class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23954–23963, 2024.

[Lu *et al.*, 2024] Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection. In *International Conference on Learning Representations*, 2024.

[Magistri *et al.*, 2024] Simone Magistri, Tomaso Trinci, Albin Soutif, Joost van de Weijer, and Andrew D. Bagdanov. Elastic feature consolidation for cold start exemplar-free

incremental learning. In *International Conference on Learning Representations*, 2024.

[Mardia and Jupp, 2009] Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

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

[Meng *et al.*, 2025] Zichong Meng, Jie Zhang, Changdi Yang, Zheng Zhan, Pu Zhao, and Yanzhi Wang. Diffclass: Diffusion-based class incremental learning. In *European Conference on Computer Vision*, pages 142–159. Springer, 2025.

[Ming *et al.*, 2023] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations*, 2023.

[Morteza and Li, 2022] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7831–7840, 2022.

[Petit *et al.*, 2023] Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3911–3920, 2023.

[Rajasegaran *et al.*, 2020] Jathushan Rajasegaran, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Mubarak Shah. itaml: An incremental task-agnostic meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13588–13597, 2020.

[Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[Rypeść *et al.*, 2023] Grzegorz Rypeść, Sebastian Cygert, Valeriya Khan, Tomasz Trzcinski, Bartosz Michał Zieliński, and Bartłomiej Twardowski. Divide and not forget: Ensemble of selectively trained experts in continual learning. In *International Conference on Learning Representations*, 2023.

[Semola *et al.*, 2024] Rudy Semola, Julio Hurtado, Vincenzo Lomonaco, and Davide Bacciu. Adaptive hyperparameter optimization for continual learning scenarios. *arXiv preprint arXiv:2403.07015*, 2024.

[Serrà *et al.*, 2018] Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018.

[Toldo and Ozay, 2022] Marco Toldo and Mete Ozay. Bring evanescent representations to life in lifelong class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16732–16741, 2022.

[Von Oswald *et al.*, 2019] Johannes Von Oswald, Christian Henning, Benjamin F Grewe, and João Sacramento. Continual learning with hypernetworks. *arXiv preprint arXiv:1906.00695*, 2019.

[Wang *et al.*, 2022] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. FOSTER: Feature boosting and compression for class-incremental learning. In *European Conference on Computer Vision*, pages 398–414. Springer, 2022.

[Wang *et al.*, 2023] Fu-Yun Wang, Da-Wei Zhou, Liu Liu, Han-Jia Ye, Yatao Bian, De-Chuan Zhan, and Peilin Zhao. BEEF: Bi-compatible class-incremental learning via energy-based expansion and fusion. In *International Conference on Learning Representations*, 2023.

[Yan *et al.*, 2021] Shipeng Yan, Jiangwei Xie, and Xuming He. DER: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021.

[You *et al.*, 2017] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[Zhou *et al.*, 2023] Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *International Conference on Learning Representations*, 2023.

[Zhu *et al.*, 2021] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021.

[Zhu *et al.*, 2023] Kai Zhu, Kecheng Zheng, Ruili Feng, Deli Zhao, Yang Cao, and Zheng-Jun Zha. Self-organizing pathway expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19204–19213, 2023.

[Zhuang *et al.*, 2022] Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. Acil: Analytic class-incremental learning with absolute memorization and privacy protection. *Advances in Neural Information Processing Systems*, 35:11602–11614, 2022.