DRL: Discriminative Representation Learning with Parallel Adapters for Class Incremental Learning

Jiawei Zhan†, Jun Liu†, Jinlong Peng, Xiaochen Chen, Bin-Bin Gao, Yong Liu, and Chengjie Wang*, *Member, IEEE*

***Abstract*—With the excellent representation capabilities of Pre-Trained Models (PTMs), remarkable progress has been made in non-rehearsal Class-Incremental Learning (CIL) research. However, it remains an extremely challenging task due to three conundrums: increasingly large model complexity, non-smooth representation shift during incremental learning and inconsistency between stage-wise sub-problem optimization and global inference. In this work, we propose the Discriminative Representation Learning (DRL) framework to specifically address these challenges. To conduct incremental learning effectively and yet efficiently, the DRL's network, called Incremental Parallel Adapter (IPA) network, is built upon a PTM and increasingly augments the model by learning a lightweight adapter with a small amount of parameter learning overhead in each incremental stage. The adapter is responsible for adapting the model to new classes, it can inherit and propagate the representation capability from the current model through parallel connection between them by a transfer gate. As a result, this design guarantees a smooth representation shift between different incremental stages. Furthermore, to alleviate inconsistency and enable comparable feature representations across incremental stages, we design the Decoupled Anchor Supervision (DAS). It decouples constraints of positive and negative samples by respectively comparing them with the virtual anchor. This decoupling promotes discriminative representation learning and aligns the feature spaces learned at different stages, thereby narrowing the gap between stage-wise local optimization over a subset of data and global inference across all classes. Extensive experiments on six benchmarks reveal that our DRL consistently outperforms other state-of-the-art methods throughout the entire CIL period while maintaining high efficiency in both training and inference phases.**

***Index Terms*—Incremental learning, classification, transformer, representation learning, catastrophic forgetting.**

I. Introduction

DEEP neural networks (DNNs) have demonstrated transformative success across numerous domains [1], [2], [3], [4], [5], typically trained via supervised or self-supervised methods [6] on static datasets like ImageNet [7]. However, these approaches struggle to adapt effectively to data stream scenarios [8], necessitating incremental learning techniques such as Class-Incremental Learning (CIL) [9], [10] and incremental object detection [11]. Among these, non-rehearsal CIL [12], [13], [14] is particularly important for privacy-sensitive applications [15], [16], [17] due to its strict prohibition against revisiting historical data. This setting intensifies the inherent *stability-plasticity dilemma* [18], inevitably leading to catastrophic forgetting [19] when acquiring new knowledge without compromising existing representations.

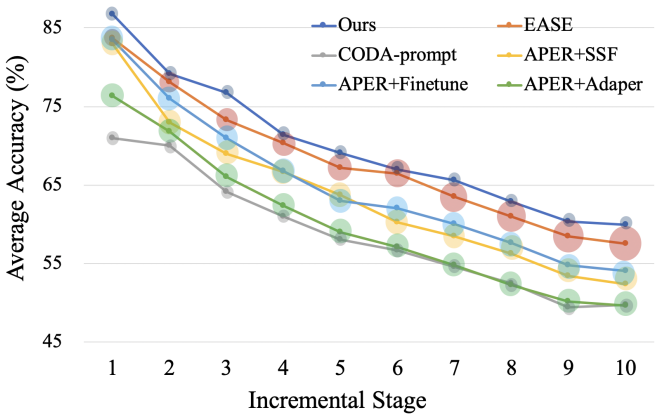


Fig. 1. Performance comparison of different class-incremental learning methods on ImageNet-A with B0 Inc20 setting, evaluated in terms of classification accuracy versus inference complexity measured by model size. The circle sizes represent the model size during inference phase, demonstrating the efficiency advantages of our DRL framework.

To address this challenge, several network-based methods have been proposed, including combining multiple networks [20] and dynamically expanding sub-networks [21], [22]. Among these, DER [23] and Progressive Networks (PN) [24] preserve previously trained models to alleviate catastrophic forgetting while expanding newly learnable models at each incremental stage to enhance plasticity. FOSTER [25] addresses the increasing model complexity in DER and PN by employing knowledge distillation to compress the model and limit its size. However, this approach requires additional training steps, increasing computational overhead.

Recent studies [26], [27], [28] have demonstrated that leveraging large Pre-Trained Models (PTMs) can significantly enhance CIL performance. Building on this insight, Zhou et al. proposed EASE [29], which achieves state-of-the-art (SOTA) performance by expanding an independent PTM with a learnable adapter [30] to acquire new knowledge. However, EASE faces increasing computational overhead during inference (see Fig. 2c) as it retains all previously trained models in memory

Manuscript received Oct 9, 2025; revised Oct 9, 2025.

J. Zhan, J. Liu, J. Peng, X.C. Chen, B.B. Gao, Y. Liu, and C.J. Wang are with Tencent Youtu Lab, Shenzhen, China. (†*Equal contributors: Jiawei Zhan and Jun Liu,* **Corresponding author: Chengjie Wang.*)

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2025.xxxxxxx.

Digital Object Identifier 10.1109/TCSVT.2025.xxxxxxx

to extract corresponding features. Furthermore, because DER, FOSTER, and EASE maintain independent models across different incremental stages, there is insufficient interaction between models from different stages (see Fig. 2a-c). This isolation prevents the current model from inheriting features from previously trained models, resulting in non-smooth representation shifts. Additionally, most CIL methods rely solely on Cross-Entropy Loss for supervision at each incremental stage, which may lead to inconsistency between stage-wise local optimization on subset data and global inference across all classes that requires more discriminative representations.

To overcome these limitations, we propose a Discriminative Representation Learning (DRL) framework that achieves superior performance with high efficiency (see Fig. 1). The DRL framework comprises two key components: an Incremental Parallel Adapter (IPA) network and a Decoupled Anchor Supervision (DAS) strategy. The IPA network builds upon a PTM and progressively enhances the model by learning a lightweight adapter at each incremental stage. This adapter inherits and propagates representation capability from the current model through parallel connections regulated by a learnable transfer gate, enabling smooth representation transitions with exceptional efficiency. To address the optimization inconsistency issue, DAS decouples constraints for positive and negative samples by comparing them with a virtual anchor: it enforces positive logits to exceed a fixed anchor k while constraining negative logits to remain below k. This decoupling promotes discriminative representation learning and aligns feature spaces across different stages, thereby narrowing the gap between stage-wise local optimization and global inference across all classes.

Extensive experiments on six benchmark datasets validate the SOTA performance of our DRL framework. On ImageNet-A, our method achieves an accuracy of 68.96%, surpassing the current SOTA by 3.62%. On VTAB and ObjectNet, we achieve accuracies of 95.73% and 72.69%, representing improvements of 2.12% and 1.85% over existing SOTA methods, respectively. The main contributions of this work are:

- We propose the novel IPA network, which achieves high training-inference efficiency and enables smooth representation transitions through a lightweight adapter and learnable transfer gate.
- We introduce DAS, a simple yet effective supervision strategy that mitigates optimization inconsistency by decoupling constraints for positive and negative samples using a fixed anchor, effectively aligning feature spaces across incremental stages. This component can be seamlessly integrated into other CIL methods.
- Our approach sets a new state-of-the-art across a comprehensive suite of six benchmarks, encompassing both standard CIL evaluations and challenging out-of-distribution benchmarks characterized by substantial domain shifts from the PTM's original training data.

The remainder of this paper is organized as follows. Section II reviews related work on CIL methods. Section III-A presents the preliminaries, while Sections III-B and III-C detail our proposed IPA network and DAS strategy, respectively. Comprehensive experimental evaluations are provided in Section IV. Finally, we discuss the limitations of our approach, suggest future research directions, and conclude with a summary of our contributions.

II. Related Work

Class Incremental Learning (CIL). CIL is essential for data streaming applications, enabling models to learn from sequentially arriving data. Methods are typically categorized based on the accessibility of training data from previous stages: *rehearsal-based CIL*, which retains a small set of old exemplars in memory to mitigate catastrophic forgetting [31], [32], [33], [34], [35], and *non-rehearsal CIL*, which fine-tunes models without relying on exemplars [36], [37], [38], [39], [40], [41], [42], [43], [44]. While rehearsal-based methods demonstrate effectiveness in preserving knowledge, storing exemplars raises significant concerns regarding data security and privacy [16], [15]. Consequently, research has increasingly focused on non-rehearsal CIL, which presents a more challenging but privacy-preserving alternative.

Network-based Methods. As a prominent category within non-rehearsal CIL [45], these methods allocate specific model parameters to each stage to address catastrophic forgetting. Recent expandable network approaches [25], [22], [46] have demonstrated particularly strong performance. However, many methods in this category (e.g., DER [23] and Progressive Networks [24]) suffer from expanding the backbone network or incorporating large modules [47], leading to increasingly cumbersome networks that lack flexibility and efficiency after multiple incremental stages. This architectural inflation poses practical limitations for long-term deployment scenarios.

Pre-Trained Model-based Methods. PTM-based methods leverage the strong representational capabilities of pre-trained models and have become a prominent research direction recently [26], [48], [49]. These approaches are generally divided into prompt-based [50], [51] and adapter-based methodologies. Recent methods such as L2P [52] and DualPrompt [53] utilize prompt tuning based on PTMs for incremental learning tasks. However, these methods typically maintain a unified prompt pool that requires continuous updating, which can lead to prompt-level forgetting and restricted representation ability over time. Alternative approaches, including APER [54] and EASE [29], expand an independent PTM with trainable adapters [55], [30] to fine-tune the model for each stage. While effective, these methods often utilize all stage models during inference, resulting in computational inefficiency and potential representation shifts due to limited interaction between models across different stages.

Supervision in Class Incremental Learning. Existing continual learning methods employ diverse supervisory signals, spanning classification losses (e.g., cross-entropy [29] and margin-based formulations like CosFace [56]) to regularization mechanisms (e.g., knowledge distillation [57], [12], [58]). Notably, softmax cross-entropy remains the predominant choice in stage-wise training paradigms across leading methods including EASE, RanPAC [59], and SAFE [60]. Nevertheless, these approaches exhibit an intrinsic limitation:

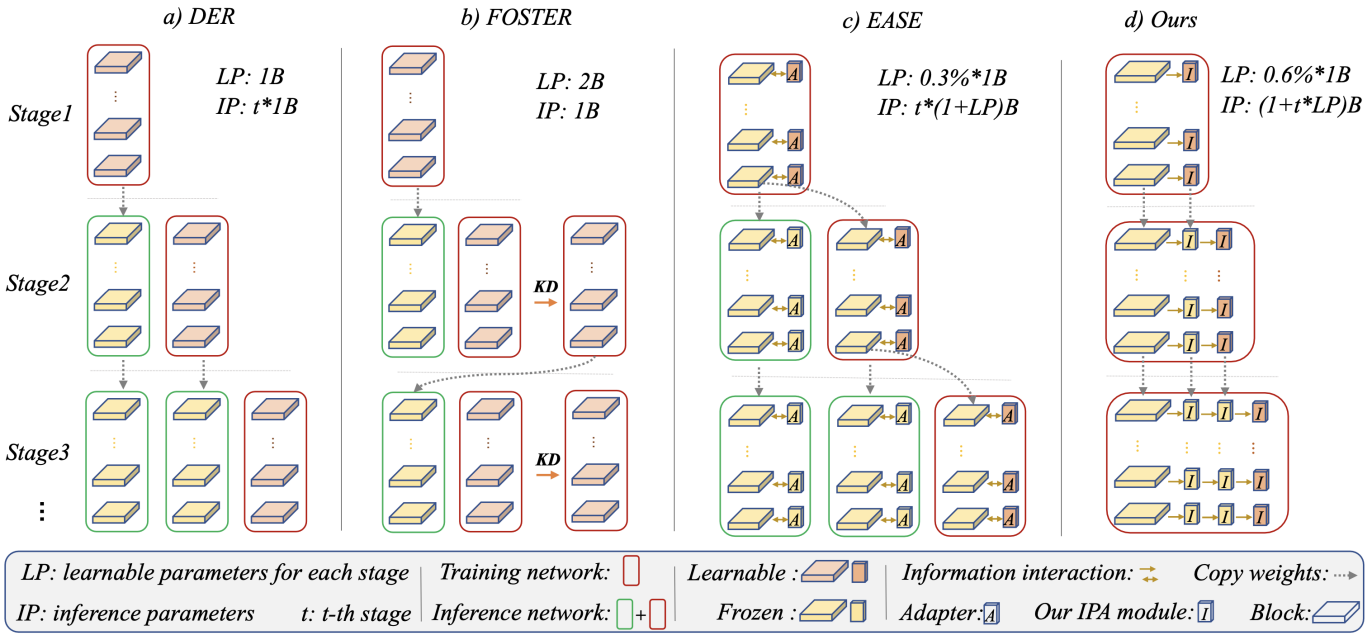


Fig. 2. **Comparison of different CIL methods.** We select three representative methods: DER, FOSTER, and EASE. '1B' denotes the total parameter of a model (e.g., ViT-B/16). a) DER and b) FOSTER face the increasingly large model complexity during the training. c) EASE faces an increasing computing overhead during inference, as well as non-smooth representation shift due to the independent training at different incremental stages. d) Our network mitigates these problems through the IPA module, which consists of a lightweight adapter with minimal parameter learning overhead at each incremental stage and a transfer gate to ensure a smooth representation shift.

inconsistent separation granularity between training objectives and inference requirements. The local optimization at each stage focuses only on the current class subset, while global inference must discriminate across all encountered classes. To resolve this misalignment, we propose Decoupled Anchor Supervision (DAS), which introduces a fixed anchor to decouple constraints on positive and negative samples. This framework simultaneously eliminates the non-independence inherent in conventional losses while ensuring consistent decision boundaries during deployment, ultimately yielding significant performance improvements across incremental learning scenarios.

III. Method

A. Preliminaries

CIL aims to train a model with training samples arriving in sequence. This incremental process can be divided into T stages. For stage $t \in \{1, 2, \cdots, T\}$, the training samples belonging to the stage t are represented as $D^t = \{X^t, Y^t\}$, where X^t is the input data, and Y^t corresponds to the associated label. The classes across different stages do not overlap, i.e., $Y^1 \cap Y^2 \cap \cdots \cap Y^T = \emptyset$. The non-rehearsal CIL satisfies $D^1 \cap D^2 \cap \cdots \cap D^T = \emptyset$, meaning no previous data is stored for replay, making this the most challenging and practical CIL scenario. During the training of the model at the t-th stage, we can only access the data D^t, while the stage identity t is not available during inference. After each stage, the trained model is evaluated on all previously seen classes, i.e., $Y^1 \cup Y^2 \cup \cdots \cup Y^t$, which requires the model to balance stability (remembering old classes) and plasticity (learning new classes) effectively.

The model of CIL can be formulated as $f(\mathbf{x}) = X \rightarrow Y$, which aims to minimize the empirical risk:

$$\sum\nolimits_{(\mathbf{x},y)\in D^1\cdots\cup D^T} L(f(\mathbf{x}), y), \tag{1}$$

Here, we decouple our model into embedding module $\Phi(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^d$ and classifier layer $\mathbf{W} \in \mathbb{R}^{d\times|Y|}$, where d represents the embedding dimension and Y represents the label space. The model output is then denoted as $f(\mathbf{x}) = \mathbf{W}^\top\Phi(\mathbf{x})$. Since our DRL is based on PTM, for the t-th stage, the embedding module can be further parameterized as $\Theta = \{\theta_t^o, \theta_t^n\}$, where θ_t^o and θ_t^n are the parameters of the previously trained model (e.g., PTM) and the new expanding network (e.g., adapter [29]), respectively. Furthermore, for the l-th transformer block [61], where $l \in \{1, \cdots, L\}$ and L represents the total number of blocks (e.g., $L = 12$ in ViT-B/16), the parameters are denoted as $\{\theta_t^{o_l}, \theta_t^{n_l}\}$. The classifier layer can be further decomposed into a combination of $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{|Y|}]$. The classifier weight for class i is $\mathbf{w}_i$, where $\mathbf{w}_i \in \mathbb{R}^{d\times 1}$.

Following the EASE [29], during the training phase, the logit for class i given by:

$$z_i = s \cdot \cos(\mathbf{w}_i, \Phi(\mathbf{x})), \tag{2}$$

where s is a learnable scale factor during the training phase. The logit z_i is passed to the softmax function to obtain the output probability:

$$p_i = \frac{e^{z_i}}{\sum_j e^{z_j}} = \frac{e^{s\cdot\cos(\mathbf{w}_i,\Phi(\mathbf{x}))}}{\sum_j e^{s\cdot\cos(\mathbf{w}_j,\Phi(\mathbf{x}))}}, \tag{3}$$

During inference, the prototype-based classifier extracts the final `[CLS]` token in ViT-B as the class center $\mathbf{c}_i$ (i.e.,

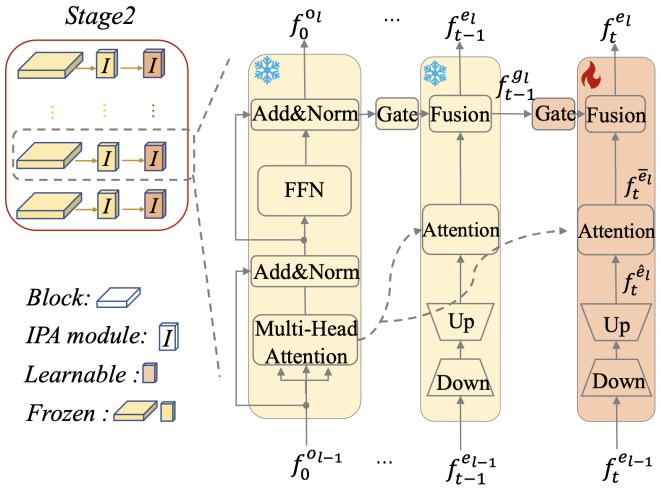


Fig. 3. **Detailed architecture of the Incremental Parallel Adapter (IPA) network.** This design incorporates a lightweight adapter structure, a reusable attention mechanism, and a transfer gate to enable seamless knowledge propagation across incremental learning stages.

prototype) for the i-th class and directly replaces $\mathbf{w}_i$. It then utilizes cosine distance to calculate the probability as:

$$\hat{p}_i = \cos(\mathbf{c}_i, \Phi(\mathbf{x})) = \frac{\mathbf{c}_i^\top \Phi(\mathbf{x})}{\|\mathbf{c}_i\|_2 \cdot \|\Phi(\mathbf{x})\|_2}. \tag{4}$$

This prototype-based inference strategy enhances robustness by leveraging class centroids that are less susceptible to the representation drift that can occur during incremental learning.

B. Incremental Parallel Adapter (IPA)

To address the large model complexity during training or inference phase [23], [25], [29] and non-smooth representation shift caused by the independent training of models at different stages, we propose Incremental Parallel Adapter (IPA) network. Inspired by PTM-based methods that demonstrate promising performance in CIL, our network is built upon a PTM and increasingly expands a IPA module in each block for current incremental stage, leading to superior performance with high efficiency. The core innovation lies in maintaining a fixed backbone while progressively integrating trainable parallel adapters, ensuring parameter efficiency while enabling continuous adaptation.

Overview of the IPA Network. The IPA network is illustrated in Fig. 2d, the details of each block are shown in Fig. 3, which mainly consists of two parts: the previously trained model parameterized by $\theta^o = \{\theta^{o_l}\}_{l=1}^L$ and the IPA module parameterized by $\theta^n = \{\theta^e, \theta^g\}$. The previously trained model θ^o is utilized to extract fundamental features, which can be either a PTM (in the first stage) or a model trained in the previous stage (the stage after the first one). This frozen shared backbone preserves established representations and prevents catastrophic forgetting. On the other hand, the IPA module consists of an efficient lightweight adapter parameterized by $\theta^e = \{\theta^{e_l}\}_{l=1}^L$ and a learnable transfer gate parameterized by $\theta^g = \{\theta^{g_l}\}_{l=1}^L$. For the l-th block during training, θ^{o_l} is fixed to retain representation ability and mitigate catastrophic forgetting while θ^{n_l} (comprising θ^{e_l} and θ^{g_l}) is learnable to acquire new knowledge. This design ensures that the model maintains stability through the fixed backbone while achieving plasticity via the adaptable IPA modules.

Lightweight Adapter in the IPA Module. The lightweight adapter consists of two 1x1 convolutional layers followed by a reusable attention module. Specifically, the input of the l-th adapter for the t-th stage is the output of the $(l-1)$-th block (e.g., $\boldsymbol{f}_t^{e_{l-1}}$ in Fig. 3). $\boldsymbol{f}_t^{e_{l-1}}$ is passed to a 1x1 convolutional layer $\mathbf{W}_{down} \in \mathbb{R}^{d\times r}$ for downsampling, followed by an activation function, and then another 1x1 convolutional layer $\mathbf{W}_{up} \in \mathbb{R}^{r\times d}$ for upsampling. The output is denoted as $\boldsymbol{f}_t^{\hat{e}_l}$. This bottleneck-like structure, consisting solely of two 1x1 convolutions, makes it extremely lightweight, adding minimal parameters while enabling feature transformation. The reusable attention module aims to enhance the correlations between features (or tokens) without introducing additional learnable parameters. Traditionally, the attention module [62] requires three additional 1x1 convolutional layers to generate the $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and utilizes $\mathbf{Q}$ and $\mathbf{K}$ to calculate the attention matrix $\mathbf{A}^e$, i.e., $\mathbf{A}^e = softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_1}})$, where d_1 is the dimension of $\mathbf{Q}$ and $\mathbf{K}$. However, this traditional approach results in additional learnable parameters. Given that the PTM inherently possesses strong representational capabilities, and the attention matrix $\mathbf{A}^o$ in the PTM encapsulates the relationships among features. We propose that the $\mathbf{A}^e$ can be replaced by the one in the PTM (i.e., $\mathbf{A}^e = \mathbf{A}^o$) without losing its plasticity. This reuse of pre-computed attention patterns significantly reduces computational overhead while maintaining expressive power. Finally, $\boldsymbol{f}_t^{\hat{e}_l}$ is treated as $\mathbf{V}$, and the output is computed as $\boldsymbol{f}_t^{\bar{e}_l} = \mathbf{A}_t^{o_l} \boldsymbol{f}_t^{\hat{e}_l}$. This attention module, which we call "reusable attention", serves as a unique form of cross-attention between the PTM and the new IPA module, effectively transferring structural knowledge from the pre-trained model to the adapter.

Transfer Gate in the IPA Module. The learnable transfer gate is responsible for connecting two parallel adapters between two adjacent stages to ensure a smooth representation shift. A naive approach would be to directly sum the old and new features. However, we have found that shallow and deep layers in the previously trained model exhibit different characteristics, and the adapter should selectively inherit knowledge from these layers. Therefore, we design a learnable transfer gate for each block to preserve essential knowledge and enhance plasticity. Specifically, the gate includes downsample and upsample layers identical to those in the lightweight adapter, followed by a sigmoid function that constrains its output to a range between 0 and 1. The input of the l-th gate for task t is denoted as $\boldsymbol{f}_{t-1}^{g_l} = (1-\gamma)\boldsymbol{f}_{t-1}^{e_l} + \gamma \boldsymbol{f}_0^{o_l}$. Here, the $\boldsymbol{f}_0^{o_l}$ represents the feature of l-th block in the PTM, and the output is a weight mask $\mathbf{M}_t^l$. Finally, we fuse the features of the previously trained network and the adapter as follows: $\boldsymbol{f}_t^{e_l} = (1-\mathbf{M}_t^l)\boldsymbol{f}_t^{\bar{e}_l} + \mathbf{M}_t^l \boldsymbol{f}_{t-1}^{g_l}$. This adaptive fusion allows the model to dynamically balance contributions from current features versus historical knowledge at a per-block granularity.

Generally, the IPA module can be integrated into all blocks. However, we have found that fusing the features of the last block reduces plasticity. Therefore, we use a transfer block to

replace the IPA module of the L-th layer and independently introduce two lightweight linear layers instead of the original Feedforward Network (FFN) in this block. This specialized handling of the final block preserves the model's capacity to learn new representations while maintaining efficiency.

During inference, performing one inference on the t-th stage model is sufficient to obtain the embedding representation $\mathbf{F}_t = [\boldsymbol{f}_0^{o_L}, \boldsymbol{f}_1^{e_L}, \cdots, \boldsymbol{f}_t^{e_L}]$. Following EASE [29], we employ the "semantic-guided prototype" to synthesize new features for old classes without accessing any old class instance and classify them using Formula 4.

C. Decoupled Anchor Supervision (DAS)

Inconsistency between stage-wise local optimization and global inference. A typical approach for training our proposed IPA model is to perform stage-wise gradient descent using the standard cross-entropy loss, where model optimization is performed relative to the class activations at the current stage. This introduces a key issue: as the logit scale drifts across training stages, the learned decision boundary becomes stage-dependent. Specifically:

- During each stage's optimization, feature activations and logits are calibrated relative to the mean activation intensity $\mu_t = \mathbb{E}_{i \in \mathcal{C}_t}[\|\mathbf{w}_i\|]$ of that stage's classes $\mathcal{C}_t$, creating stage-specific reference points.
- During global inference, features from different stages must be compared directly despite lacking a consistent reference point. This incompatibility results in misaligned and suboptimal decision boundaries

This reference point inconsistency prevents meaningful comparison of logits across stages, undermining the model's ability to maintain discriminability across all learned classes. To alleviate this problem, we propose the Decoupled Anchor Supervision (DAS) to optimize the representation learning and enhance class cross-stage compatibility.

Specifically, DAS introduces an predefined virtual anchor k that serves as a fixed reference point across all incremental stages and decouples constraints for positive and negative samples, enabling consistent logit interpretation and feature space alignment. The core objective of DAS is to enforce the following anchor-based constraints:

$$\text{Positive: } z_i > k, \tag{5}$$

$$\text{Negative: } z_j < k \quad \text{for } j \neq i. \tag{6}$$

where $z_i = s \cdot \cos(\mathbf{w}_i, \Phi(\mathbf{x}))$ is the logit for class i.

To achieve these constraints, we design the following decoupled supervision signals:

$$L^{pos} = -\sum\nolimits_i y_i \log(p_i^{pos}), \quad L^{neg} = -\log(p^{neg}), \tag{7}$$

Here, y_i is the groundtruth label, and the p_i^{pos} and p^{neg} calculate as follows:

$$p_i^{pos} = \frac{e^{z_i}}{e^{z_i} + e^k} = \frac{e^{s \cdot \cos(\mathbf{w}_i, \Phi(\mathbf{x}))}}{e^{s \cdot \cos(\mathbf{w}_i, \Phi(\mathbf{x}))} + e^k}, \tag{8}$$

$$p^{neg} = \frac{e^k}{\sum_{j,j \neq i}^{C} e^{z_j} + e^k} = \frac{e^k}{\sum_{j,j \neq i} e^{s \cdot \cos(\mathbf{w}_j, \Phi(\mathbf{x}))} + e^k}. \tag{9}$$

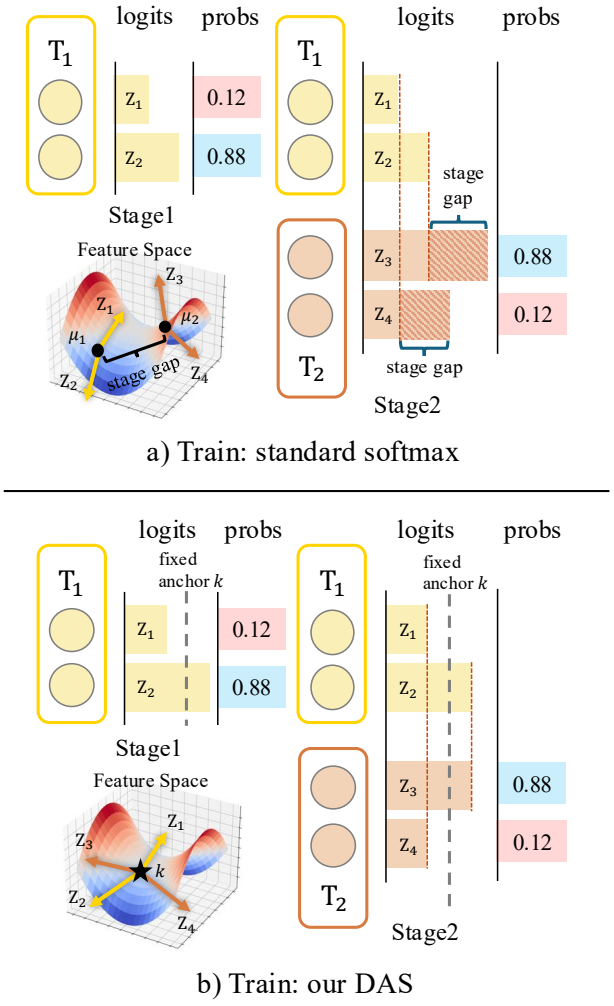


Fig. 4. **Comparisons of decision boundaries.** (a) Standard softmax exhibits inconsistent separation boundaries across incremental stages; (b) Our proposed DAS establishes consistent decision boundaries through a fixed virtual anchor, enabling stable representation learning across all incremental process.

Additionally, most CIL tasks involve single-label classification. Therefore, we can simplify Equation 7 for single-label classification tasks as follows:

$$L^{pos} = -\log(p^{pos}), \quad L^{neg} = -\log(p^{neg}), \tag{10}$$

To balance L^{pos} and L^{neg}, we set λ_p, λ_n as the loss weight, and our proposed DAS is defined as:

$$L_{das} = \lambda_p L^{pos} + \lambda_n L^{neg}. \tag{11}$$

Analysis of DAS. Fig. 4 illustrates a fundamental limitation of standard softmax-based training in incremental learning and its resolution via Decoupled Anchor Supervision (DAS). In standard softmax training (Fig. 4a, upper), classifiers for sequential tasks are optimized independently. While this approach achieves local discriminability within each individual task (e.g., within Task 1, the predicted class is determined if its logit exceeds the logits of all other classes within Task 1), it suffers from decision boundary drift due to semantic shifts and

TABLE I
COMPARISON OF AVERAGE AND FINAL ACCURACY ACROSS SIX BENCHMARK DATASETS. ALL METHODS, IMPLEMENTED WITHOUT EXEMPLARS AND USING THE VIT-B/16-IN21K PRETRAINED WEIGHTS, ARE FAIRLY COMPARED ON IMAGENET-R/A (IN-R/A) AND OTHER BENCHMARKS. THE BEST PERFORMANCES ARE HIGHLIGHTED IN **BOLD**.

Methods	CIFAR B0 Inc5		IN-R B0 Inc5		IN-A B0 Inc20		ObjNet B0 Inc10		Omni B0 Inc30		VTAB B0 Inc10	
	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
Finetune	38.90	20.17	21.61	10.79	24.28	14.51	19.14	8.73	23.61	10.57	34.95	21.25
Finetune Adapter [63]	60.51	49.32	47.59	40.28	45.41	41.10	50.22	35.95	62.32	50.53	48.91	45.12
LwF [12]	46.29	41.07	39.93	26.47	37.75	26.84	33.01	20.65	47.14	33.95	40.48	27.54
SDC [40]	68.21	63.05	52.17	49.20	29.11	26.63	39.04	29.06	60.94	50.28	45.06	22.50
L2P [52]	85.94	79.93	66.53	59.22	49.39	41.71	63.78	52.19	73.36	64.69	77.11	77.10
DualPrompt [53]	87.87	81.15	63.31	55.22	53.71	41.67	59.27	49.33	73.92	65.52	83.36	81.23
CODA-Prompt [50]	89.11	81.96	64.42	55.08	53.54	42.73	66.07	53.29	77.03	68.09	83.90	83.02
SimpleCIL [54]	87.57	81.26	62.58	54.55	59.77	48.91	65.45	53.59	79.34	73.15	85.99	84.38
APER w/ Adapter [54]	90.65	85.15	72.35	64.33	60.47	49.37	67.18	55.24	80.75	74.37	85.95	84.35
EASE [29]	91.51	85.80	78.31	70.58	65.34	55.04	70.84	57.86	81.11	74.85	93.61	93.55
DRL	**92.01**	**86.91**	**78.87**	**72.20**	**68.96**	**59.38**	**72.69**	**60.29**	**81.26**	**74.98**	**95.73**	**95.01**

feature distribution misalignment across tasks. For instance, the boundary separating z_3 and z_4 in Task 2 drifts relative to the boundary established for z_1 and z_2 in Task 1 (Fig. 4a, upper). This drift introduces task-specific biases into the feature/logit spaces, undermining global discriminability and distorting distance-based metrics (e.g., prototype matching). DAS (Fig. 4b, lower) addresses this by anchoring supervision to a fixed reference point k. For a sample $\mathbf{x}_1$ from class 1 with feature $\mathbf{f}_{x_1} = \Phi(\mathbf{x}_1)$, DAS decouples classification by requiring that the positive logit $\mathbf{w}_1^\top \mathbf{f}_{x_1}$ exceeds the anchor k, while simultaneously enforcing that all negative logits $\mathbf{w}_j^\top \mathbf{f}_{x_1}$ $(j \neq 1)$ fall below this same anchor value. Crucially, k is task-agnostic and fixed (e.g., $k = 0$), establishing an absolute reference axis that disentangles supervision from task-specific contexts. This unified anchoring ensures all task-specific classifiers ($\mathbf{w}_i$) and features are implicitly aligned to the same global reference, guaranteeing cross-task compatibility in both logit and feature spaces (Fig. 4b), thus stabilizing representation learning (Table IV). Besides, DAS naturally accommodates discriminative margins: replacing k with *separate anchors* k_+ and k_- introduces a margin $\delta = k_+ - k_- > 0$. This can create a "buffer zone" that further enhances feature separability.

Given feature distributions inherited from pretrained models exhibit advantageous generalization properties, we incorporate knowledge distillation to constrain feature space during incremental updates. The final optimization objective integrates DAS with distillation regularization:

$$L_{total} = L_{das} + \alpha L_{kd}. \tag{12}$$

Here, α is the loss weight for L_{kd}, L_{kd} is the loss of the final embedding (such as the final `[CLS]` token in ViT) between the PTM and the newly inserted IPA module. For simplicity, we use cosine distance as the metric for L_{kd}.

IV. EXPERIMENTS

In this section, we comprehensively evaluate DRL against state-of-the-art methods on six benchmark datasets, utilizing diverse pre-trained models and data splits. This demonstrates DRL's superior performance. Ablation studies confirm its effectiveness, and visual analyses provide additional insights.

TABLE II
COMPARISON OF AVERAGE AND FINAL ACCURACY ACROSS TWO BENCHMARK DATASETS. ALL METHODS, IMPLEMENTED WITHOUT EXEMPLARS AND USING THE VIT-B/16-IN1K PRETRAINED WEIGHTS, ARE FAIRLY COMPARED ON IMAGENET-R/A (IN-R/A) AND OTHER BENCHMARKS. THE BEST PERFORMANCES ARE HIGHLIGHTED IN **BOLD**.

Methods	ObjNet B0 Inc20		ImageNet-A B0 Inc20	
	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
iCaRL [13]	33.43	19.18	29.22	16.16
LUCIR [64]	41.17	25.89	31.09	18.59
DER [23]	35.47	23.19	33.85	22.27
FOSTER [25]	37.83	25.07	34.82	23.01
MEMO [65]	38.52	25.41	36.37	24.46
FACT [66]	60.59	50.96	60.13	49.82
SimpleCIL [54]	62.11	51.13	59.67	49.44
APER w/ SSF [54]	68.75	56.79	63.59	52.67
EASE [29]	70.44	58.37	65.74	57.28
DRL	**72.81**	**61.00**	**69.42**	**59.97**

Datasets. We evaluate the performance on six datasets: CIFAR100 [67], ImageNet-R [68], ImageNet-A [69], ObjectNet [70], OmniBench [71], and VTAB [72]. This selection encompasses both standard class-incremental learning benchmarks (CIFAR100, ImageNet-R) and challenging out-of-distribution datasets (ImageNet-A, ObjectNet, OmniBench, VTAB) that exhibit significant domain shifts from the pre-training data. The datasets vary in scale and complexity: CIFAR100 contains 100 classes, ImageNet-R and ImageNet-A each have 200 classes, ObjectNet comprises 200 classes, OmniBench includes 300 classes, and VTAB consists of 50 classes. For ablation studies and detailed analysis, we primarily focus on ImageNet-A (known for its challenging samples that challenge ImageNet pre-trained models) and VTAB (featuring diverse classes from multiple complex domains), as they effectively showcase the robustness of our approach under distribution shifts. Following established benchmark protocols [13], [52], [73], we employ the 'B-m Inc-n' class

TABLE III
ABLATION STUDY ON THE COMPONENTS OF DECOUPLED ANCHOR SUPERVISION (DAS). THE EVALUATION, PERFORMED ON IMAGENET-A AND VTAB USING VIT-B/16-IN21K WEIGHTS, COMPARES DIFFERENT LOSS CONFIGURATIONS INCLUDING CROSS-ENTROPY (L_{ce}), KNOWLEDGE DISTILLATION (L_{kd}), AND DAS (L_{das}), DEMONSTRATING THE COMPLEMENTARY BENEFITS OF EACH COMPONENT AND THE OPTIMAL COMBINATION IN DRL.

Methods	Components			IN-A B0 Inc20		VTAB B0 Inc10	
	L_{ce}	L_{kd}	L_{das}	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
EASE [29]	✓	✗	✗	65.34	55.04	93.61	93.55
EASE [29]	✓	✓	✗	66.51	55.50	91.64	90.99
DRL	✓	✗	✗	66.16	56.05	94.48	93.83
DRL	✓	✓	✗	67.24	57.12	94.72	94.03
DRL	✗	✓	✓	**68.96**	**59.38**	**95.73**	**95.01**

TABLE IV
COMPARATIVE STUDY OF DIFFERENT LOSS FUNCTIONS INTEGRATED WITH OUR DRL FRAMEWORK. THE EVALUATION, PERFORMED ON IMAGENET-A AND VTAB USING VIT-B/16-IN21K WEIGHS, INCLUDES BINARY CROSS-ENTROPY (L_{bce}), COSFACE (L_{cf}), AND DECOUPLED ANCHOR SUPERVISION (L_{das}), HIGHLIGHTING DAS'S SUPERIOR PERFORMANCE IN ENHANCING FEATURE DISCRIMINABILITY AND CROSS-STAGE COMPATIBILITY.

Components			IN-A B0 Inc20		VTAB B0 Inc10	
L_{bce}	L_{cf}	L_{das}	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
✓	✗	✗	63.04	53.13	94.25	93.28
✗	✓	✗	67.34	57.25	95.02	94.21
✓	✓	✗	67.30	57.13	94.65	93.64
✗	✗	✓	**68.96**	**59.38**	**95.73**	**95.01**

split convention, where m denotes the number of classes in the initial stage and n represents the number of classes introduced in each incremental stage.

Evaluation Metric. Consistent with standard evaluation practices in class-incremental learning [13], we define $\mathcal{A}_t$ as the Top-1 accuracy after the t-th incremental stage. Our primary evaluation metrics are: (1) $\mathcal{A}_T$, representing the final performance after all incremental stages, which reflects the model's ability to retain knowledge across the entire learning trajectory; and (2) $\bar{\mathcal{A}} = \frac{1}{T}\Sigma_{t=1}^{T}\mathcal{A}_t$, the average performance across all stages, which comprehensively captures the stability-plasticity trade-off throughout the incremental learning process.

Comparison Methods. We compare DRL framework against state-of-the-art methods from two categories: (1) PTM-based CIL methods including L2P [52], DualPrompt [53], CODA-Prompt [50], APER [54], and EASE [29], which leverage pre-trained vision transformers; and (2) Conventional CIL approaches such as LwF [12], SDC [40], iCaRL [13], LUCIR [64], DER [23], FOSTER [25], MEMO [65] and FACT [66], which represent established techniques in incremental learning. All methods are initialized with identical pre-trained models to ensure fair comparison.

Training Details. Following established protocols in [52], [54], we utilize Vision Transformer Base (ViT-B/16) architectures pre-trained on ImageNet-21K (IN21K) and ImageNet-1K (IN1K). For DRL, optimization is performed using SGD [74] with momentum (0.9), weight decay (5×10^{-4}), and batch size 48, trained for 20 epochs per incremental stage. The learning rate follows a cosine annealing schedule starting from 0.01. Key hyperparameters are set as: $\alpha = 0.5$ (loss balancing), $\lambda_p = 3$ (positive anchor weighting), $\lambda_n = 1$ (negative anchor weighting), and $\gamma = 0.9$ (feature retention ratio). The virtual anchor mechanism uses $k = 1$ neighbor, while the dimensionality reduction in adapter layers employs $r = 48$ dimensions for both $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d\times48}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{48\times d}$. Within the final transformer block, the IPA module implements sequential linear transformations with $\mathbf{W}_{\text{first}} \in \mathbb{R}^{768\times768}$ and $\mathbf{W}_{\text{second}} \in \mathbb{R}^{768\times768}$. Unless otherwise specified, all experiments utilize ViT-B/16-IN21K as the pre-trained model for initialization to ensure consistent and fair comparisons across different benchmark evaluations.

TABLE V
GENERALIZATION STUDY OF DAS ACROSS FIVE STATE-OF-THE-ART CIL METHODS. THE EVALUATION, PERFORMED ON IMAGENET-A AND VTAB USING VIT-B/16-IN21K WEIGHS, VALIDATES THAT REPLACING THE CONVENTIONAL CROSS-ENTROPY LOSS WITH DAS YIELDS CONSISTENT AND SIGNIFICANT PERFORMANCE GAINS.

Methods	Components		IN-A B0 Inc20		VTAB B0 Inc10	
	L_{ce}	L_{das}	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
ESN [75]	✓	✗	52.66	41.54	86.34	69.23
ESN [75]	✗	✓	**53.94**	**42.92**	**88.77**	**72.61**
RanPAC [59]	✓	✗	70.35	62.17	92.30	91.92
RanPAC [59]	✗	✓	**71.01**	**63.06**	**93.52**	**92.51**
APER [54]	✓	✗	64.63	53.85	90.20	86.16
APER [54]	✗	✓	**65.54**	**54.25**	**92.57**	**88.84**
EASE [29]	✓	✗	65.34	55.04	93.61	93.55
EASE [29]	✗	✓	**67.71**	**58.85**	**95.36**	**94.57**
SAFE [60]	✓	✗	73.14	65.90	95.22	94.91
SAFE [60]	✗	✓	**73.95**	**66.84**	**95.96**	**95.43**

A. Comparisons

This section presents a comprehensive comparison of DRL with other state-of-the-art CIL methods using ViT-B/16-IN21K or ViT-B/16-IN1K weights on benchmark datasets.

Table I presents a comparison of DRL against state-of-the-art methods using ViT-B/16-IN21K across six benchmark datasets. DRL consistently outperforms all competing methods, demonstrating superior performance on both standard CIL benchmarks and challenging out-of-distribution datasets. Notably, DRL achieves significant improvements over current state-of-the-art methods including EASE, APER, and DualPrompt across all settings.

On out-of-distribution datasets where domain shift presents substantial challenges, DRL shows approximately 2-4% improvement over the strongest baseline (EASE). Specifically, on ImageNet-A, VTAB, and ObjectNet, DRL achieves $\bar{\mathcal{A}}$ scores of 68.96%, 95.73%, and 72.69%, outperforming EASE by 3.62%, 2.12%, and 1.85% respectively. These results highlight DRL's exceptional capability to handle distribution shifts while maintaining incremental learning performance.

To further validate the generalization capability across

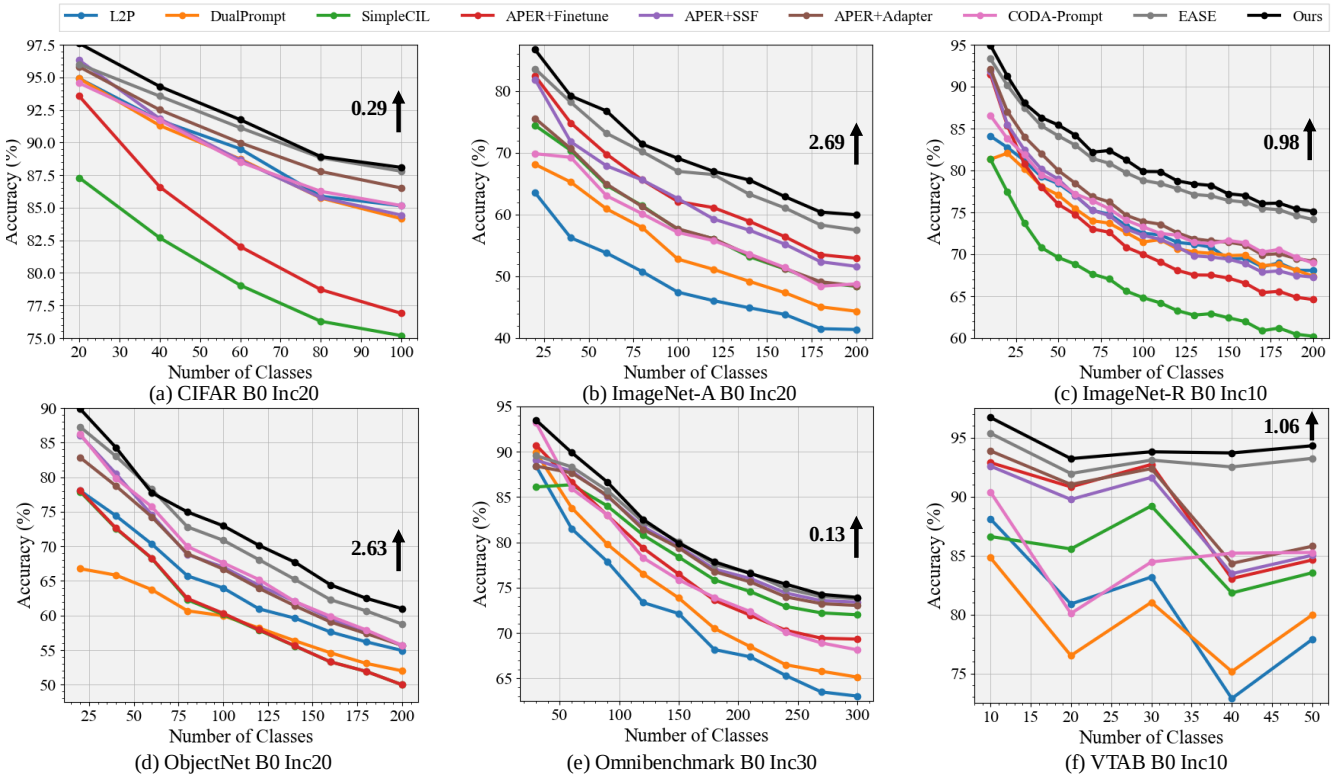


Fig. 5. **Incremental learning performance progression across multiple datasets using ViT-B/16-IN1K weights.** The percentage values indicate DRL's relative improvement over the second-best method at the final task, demonstrating consistent superiority throughout the learning trajectory.

different architectures and training configurations, we also conduct experiments following APER and EASE using ViT-B/16-IN1K as the pre-trained model under various data splits. As shown in Table II and Fig. 5, DRL maintains competitive performance across all settings. It notably outperforms the second-best method by 2.37% on ObjectNet and 3.68% on ImageNet-A, while also showing improvements of 2.69%, 1.06%, and 2.63% on ImageNet-A, VTAB, and ObjectNet respectively compared to EASE. These results consistently demonstrate the robustness and generalizability of our method under diverse experimental conditions.

B. Ablation Study

In this section, we conduct systematic ablation studies to investigate the contribution of each component in DRL, organized around four key aspects: the Decoupled Anchor Supervision mechanism, architectural design choices, hyper-parameter sensitivity, and computational efficiency.

Effect of Decoupled Anchor Supervision (DAS). We first evaluate the effectiveness of our proposed DAS module in Table III. Here, EASE serves as our baseline, and L_{ce} denotes the standard Cross-Entropy loss. Since the total loss of DRL incorporates the knowledge distillation loss L_{kd}, we conducted an additional experiment with our baseline to ensure a fair comparison. The results clearly indicate that the proposed DAS is highly effective. In particular, the configuration using DRL without L_{ce} but with L_{kd} and L_{das} achieves the best performance, with $\bar{\mathcal{A}}$ scores of 68.96% on ImageNet-A, outperforming the setting with only L_{ce} and L_{kd} by 1.72%. This demonstrates that DAS not only complements but can also effectively replace the conventional cross-entropy loss, leading to robust feature learning in incremental settings.

We further compare DAS against other popular loss functions to highlight its distinctive advantages. As shown in Table IV, we consider Binary Cross-Entropy (L_{bce}) and CosFace loss (L_{cf}) as competitive alternatives. While combinations of these losses yield reasonable results, DAS alone achieves superior performance across both ImageNet-A and VTAB benchmarks. Specifically, on ImageNet-A, DAS surpasses the best competing loss combination by 1.62% in terms of $\bar{\mathcal{A}}$, suggesting that its decoupled design provides a more suitable optimization objective for handling the stability-plasticity trade-off in incremental learning.

To validate the general applicability of DAS, we integrate it into five state-of-the-art incremental learning methods: APER, ESN, EASE, RanPAC, and SAFE. The consistent improvements observed in Table V demonstrate DAS's versatility. Across all methods, simply replacing the conventional cross-entropy loss with DAS yields significant performance gains, with improvements ranging from 0.66% to 2.37% in $\bar{\mathcal{A}}$ on ImageNet-A. This plug-and-play compatibility, coupled with the absence of additional computational overhead, positions DAS as a valuable component for enhancing existing continual learning frameworks.

Effect of Reusable Attention in IPA. We then examine the importance of attention mechanisms within the lightweight

TABLE VI
ARCHITECTURAL COMPARISON OF VARIOUS ATTENTION STRATEGIES WITHIN THE LIGHTWEIGHT ADAPTER COMPONENT OF IPA NETWORK. THE EVALUATION INCLUDES NO ATTENTION (`n-att`), STANDARD SELF-ATTENTION (`s-att`), AND REUSABLE ATTENTION (`r-att`), DEMONSTRATING THE EFFECTIVENESS OF REUSING PRE-TRAINED ATTENTION PATTERNS FOR EFFICIENT KNOWLEDGE TRANSFER.

Methods	ImageNet-A B0 Inc20		VTAB B0 Inc10	
	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
`n-att`	67.90	57.59	93.96	93.36
`s-att`	68.78	**59.76**	95.57	**95.14**
`r-att`	**68.96**	59.38	**95.73**	95.01

TABLE VII
ARCHITECTURAL COMPARISON OF DIFFERENT TRANSFER GATE DESIGNS IN IPA NETWORK. THE EVALUATION INCLUDES DIRECT SUMMATION (`sum`), PARTIAL GATING (`gate/part`), ADAPTIVE GATING (`gate/adapt`), AND EXTENDED FUSION (`gate/extra`), DEMONSTRATING THE OPTIMAL BALANCE OF PERFORMANCE AND EFFICIENCY ACHIEVED BY OUR ADAPTIVE MECHANISM.

Methods	ImageNet-A B0 Inc20		VTAB B0 Inc10	
	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
`sum`	66.53	56.84	94.41	93.64
`gate/part`	67.56	57.61	94.94	93.81
`gate/adapt`	68.96	59.38	95.73	95.01
`gate/extra`	69.21	59.62	95.80	95.13

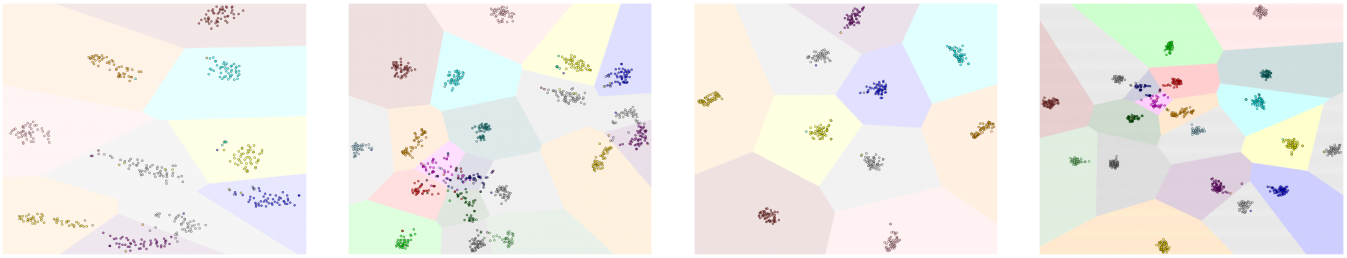
adapter component of IPA. Table VI compares three configurations: `n-att` (no attention modules), `s-att` (standard self-attention), and our proposed `r-att` (reusable attention). The empirical results demonstrate that incorporating attention substantially enhances model performance, with both `s-att` and `r-att` outperforming the attention-free baseline by approximately 1-1.5% in $\bar{\mathcal{A}}$ on ImageNet-A. The key innovation of our reusable attention mechanism lies in its ability to leverage the pre-trained model's original attention matrix $\mathbf{A}^o$, which inherently captures rich feature interdependencies learned during large-scale pre-training. This approach provides two significant advantages: first, it preserves the model's plasticity by maintaining the semantic relationships encoded in the pre-trained attention patterns; second, it eliminates additional parametric overhead since no new attention parameters need to be learned. The performance gap between `r-att` (68.96%) and `s-att` (68.78%) on ImageNet-A, though modest, confirms that distilled attention knowledge can be effectively transferred to downstream adaptation modules while maintaining efficiency. This architectural innovation demonstrates that careful reuse of pre-existing components can enhance functionality without compromising computational efficiency.

Effect of Transfer Gate in IPA. We further investigate different feature fusion strategies for the transfer gate, which plays a critical role in balancing current and historical information flow. Table VII compares four approaches: direct summation (`sum`) defined as $\boldsymbol{f}_t^{e_l} = \boldsymbol{f}_t^{\bar{e}_l} + \boldsymbol{f}_{t-1}^{g_l}$; partial gating (`gate/part`) formulated as $\boldsymbol{f}_t^{e_l} = \boldsymbol{f}_t^{\bar{e}_l} + \mathbf{M}_t^l \boldsymbol{f}_{t-1}^{g_l}$; adaptive gating (`gate/adapt`) expressed as $\boldsymbol{f}_t^{e_l} = (1 - \mathbf{M}_t^l)\boldsymbol{f}_t^{\bar{e}_l} + \mathbf{M}_t^l \boldsymbol{f}_{t-1}^{g_l}$; and extended fusion (`gate/extra`) computed as $\boldsymbol{f}_t^{e_l} = \frac{1}{2}\left[(2 - \mathbf{M}_t^l - \mathbf{M}_0^l)\boldsymbol{f}_t^{\bar{e}_l} + \mathbf{M}_t^l \boldsymbol{f}_{t-1}^{e_l} + \mathbf{M}_0^l \boldsymbol{f}_0^{o_l}\right]$. Our systematic exploration establishes the adaptive gating mechanism (`gate/adapt`) as the optimal solution, delivering superior performance without parameter overhead. As Table VII demonstrates, it achieves significant improvements over alternatives: +2.43% $\bar{\mathcal{A}}$ versus direct summation (`sum`) and +1.40% over partial gating (`gate/part`) on ImageNet-A, while maintaining full parameter efficiency. The extended fusion variant (`gate/extra`) yields only marginal gains (0.25% $\bar{\mathcal{A}}$ improvement) at substantial cost—introducing 0.29% additional parameters and increased architectural complexity. Crucially, our adaptive gating solution provides three decisive advantages: superior parameter efficiency requiring zero additional parameters beyond the base architecture; optimal feature integration balancing historical and current knowledge; and enhanced stability-plasticity tradeoff through dynamic feature weighting. The 2.43% performance leap over direct summation validates our core design principle: selective feature integration through learnable masks is fundamental for effective incremental learning. While future work may explore hierarchical formulations like $\boldsymbol{f}_t^{e_l} = \frac{1}{t+1}[(t + 1 - \sum_{i=1}^{t-1} \mathbf{M}_i^l - \mathbf{M}_0^l)\boldsymbol{f}_t^{\bar{e}_l} + \sum_{i=1}^{t-1} \mathbf{M}_i^l \boldsymbol{f}_i^{e_l} + \mathbf{M}_0^l \boldsymbol{f}_0^{o_l}]$, these inevitably increase complexity. Our adaptive gating remains the preferred solution, achieving the optimal equilibrium between accuracy (95.73% $\bar{\mathcal{A}}$ on VTAB), efficiency, and architectural elegance.

TABLE VIII
ARCHITECTURAL ANALYSIS OF BOTTLENECK WIDTHS IN THE FINAL TRANSFORMER BLOCK OF IPA NETWORK. THE EVALUATION COMPARES VARIOUS BOTTLENECK CONFIGURATIONS (`768→r→768`) WITH DIFFERENT HIDDEN DIMENSIONS `r`, DEMONSTRATING THE TRADE-OFF BETWEEN MODEL COMPLEXITY AND REPRESENTATION CAPACITY IN ADAPTER-BASED CONTINUAL LEARNING.

Architecture	ImageNet-A B0 Inc20		VTAB B0 Inc10	
	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
`768→192→768`	68.81	59.03	95.52	94.88
`768→384→768`	68.78	59.04	95.51	94.92
`768→768→768`	**68.96**	59.38	**95.73**	95.01
`768→1536→768`	68.98	59.24	95.63	**95.14**
`768→2304→768`	68.98	**59.83**	95.56	94.94
`768→3072→768`	68.97	59.82	95.61	94.99

Architectural Optimization in Final Block. We analyze the design choices for the linear transformations within the IPA network's final block, where we replace the original Feed-forward Network (FFN) with two lightweight linear layers. Table VIII compares various dimensionality configurations, where `768→384→768` indicates a bottleneck architecture with hidden dimension 384, and the original FFN is represented as `768→3072→768` (standard ViT expansion factor of 4). Our experiments reveal that the configuration `768→768→768` achieves an optimal balance between performance and parameter efficiency. While larger hidden dimensions (`1536`, `2304`, `3072`) show marginal improvements in certain metrics, the gains are insufficient to justify the substantial parameter increase. For instance, expanding to `768→3072→768` increases parameters by 4× but does not improves $\bar{\mathcal{A}}$ compared to our chosen configuration. This suggests that for adapter-



(a) $T = 1$ w/o DAS. (b) $T = 2$ w/o DAS. (c) $T = 1$ with DAS. (d) $T = 2$ with DAS.

Fig. 6. **t-SNE visualizations of feature representations on VTAB dataset with B0 Inc10 setting,** comparing the effect of Decoupled Anchor Supervision (DAS) on feature separability and cluster compactness across incremental learning stages.

TABLE IX
SENSITIVITY ANALYSIS OF LOSS WEIGHTING PARAMETERS IN OUR DRL FRAMEWORK. THE STUDY SYSTEMATICALLY ASSESSES THE KNOWLEDGE DISTILLATION WEIGHT (α), POSITIVE CONSTRAINT WEIGHT (λ_p), AND NEGATIVE CONSTRAINT WEIGHT (λ_n), REVEALING ROBUST PERFORMANCE ACROSS A WIDE RANGE OF VALUES AND IDENTIFYING THE OPTIMAL CONFIGURATION FOR TASK PERFORMANCE.

α	λ_p	λ_n	ImageNet-A B0 Inc20		VTAB B0 Inc20	
			$\bar{\mathcal{A}}$	$\mathcal{A}_T$	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
0	1	2	67.29	57.01	94.90	93.97
0.5	1	2	**68.96**	**59.38**	**95.73**	**95.01**
1	1	2	68.74	58.72	95.21	94.30
3	1	2	68.11	57.47	94.73	93.72
5	1	2	67.51	56.35	94.17	93.22
0.5	1	0.5	65.52	54.97	94.65	93.75
0.5	1	1	68.46	58.07	95.00	94.16
0.5	1	3	68.17	58.06	95.40	94.49
0.5	1	5	67.16	57.47	94.93	94.16
0.5	0.5	1	66.72	56.93	94.84	94.35
0.5	1	1	68.46	58.07	95.00	94.16
0.5	3	1	68.91	59.06	95.62	94.97
0.5	5	1	67.54	56.38	94.84	94.07

based continual learning, maintaining dimensional consistency with the pre-trained features (768 dimensions) provides sufficient representational capacity without introducing unnecessary complexity. This optimization strategy reduces model complexity compared to the standard FFN while maintaining competitive performance, demonstrating that efficient adapter design is crucial for practical continual learning systems.

Hyperparameter Sensitivity. We conduct extensive sensitivity analysis on key hyperparameters to assess model robustness. Table X examines the impact of virtual anchor parameter k, showing stable performance within the range [0,2]. It is important to clarify that in our current implementation, we employ a single shared anchor k for both positive and negative constraints, as defined by $z_i > k$ for positive samples and $z_j < -k$ for negative samples. This symmetric design provides a balanced separation boundary while maintaining simplicity.

When $k = 0$, the anchor margin becomes zero, providing the most direct form of decoupled supervision where positive and negative constraints are completely separated without additional margin. At $k = 1$, the predefined virtual anchor introduces a positive margin that further enhances feature discriminability by creating clearer separation boundaries between classes. This margin acts as a fixed reference point across all incremental stages, enforcing consistent logit interpretation and feature space alignment. The performance improvement from $k = 0$ to $k = 1$ confirms that the margin provides additional benefits for feature separation beyond the basic decoupling mechanism. While DAS can be extended to use separate anchors k_+ and k_- for positive and negative constraints respectively, our experiments focus on the symmetric case for simplicity and stability.

However, excessively large k values ($\geq$3) gradually degrade performance, as overly aggressive margin settings may impede intra-class cohesion by introducing excessive separation requirements. The supervision signal becomes too stringent, making it difficult to maintain compact class distributions while satisfying the margin constraints. This demonstrates the importance of maintaining an appropriate balance between leveraging margin-based separation and preserving learnable class distributions.

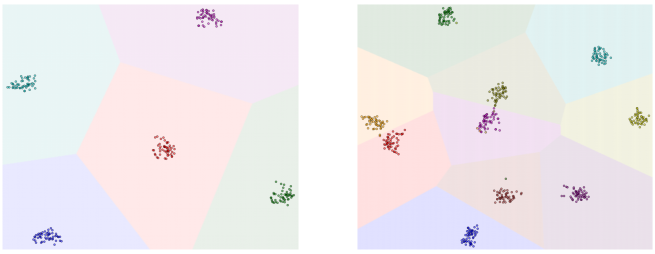
Similarly, Table IX analyzes the sensitivity to loss weighting parameters α, λ_p, and λ_n. The results indicate robust performance across a wide range of values, with optimal performance achieved at $\alpha = 0.5$, $\lambda_p = 3$, $\lambda_n = 1$. The model shows particular stability for $\lambda_p, \lambda_n \in [1, 3]$, with performance variations of less than 0.5% within this range. This robustness is advantageous for practical applications where extensive hyperparameter tuning may be impractical. The sensitivity analysis collectively demonstrates that DRL provides consistent performance across reasonable parameter choices, reducing the burden of meticulous hyperparameter optimization while maintaining competitive performance through its well-designed architectural components.

C. Extended Analysis

Parameter and Computational Efficiency. To illustrate the efficiency of DRL, we provide a comprehensive analysis of parameter counts and computational requirements during both training and inference phases, as summarized in Fig. 6. The baseline ViT-B/16 model contains approximately 1B parameters (denoted as '1B'). Our DRL approach introduces only 0.6% additional trainable parameters relative to the base model, demonstrating exceptional parameter efficiency. During

TABLE X
SENSITIVITY ANALYSIS OF VIRTUAL ANCHOR PARAMETER k IN DECOUPLED ANCHOR SUPERVISION. THE EVALUATION REVEALS PERFORMANCE STABILITY ACROSS A WIDE RANGE OF ANCHOR VALUES, HIGHLIGHTING THE IMPORTANCE OF APPROPRIATE MARGIN SETTINGS FOR EFFECTIVE FEATURE SEPARATION AND CROSS-STAGE ALIGNMENT.

k	IN-A B0 Inc20			VTAB B0 Inc10		
	s	$\bar{\mathcal{A}}$	$\mathcal{A}_T$	s	$\bar{\mathcal{A}}$	$\mathcal{A}_T$
0.0	12.47	68.67	58.53	10.11	94.96	94.08
0.5	13.75	68.81	58.72	11.29	95.21	94.47
1.0	14.95	**68.96**	**59.38**	12.46	**95.73**	**95.01**
2.0	17.33	68.58	59.18	14.77	94.99	94.23
3.0	19.53	67.85	57.60	17.00	94.35	93.51
5.0	23.60	65.16	55.83	21.49	94.05	93.04



(a) Stage 1 feature embeddings on CIFAR100

(b) Stage 2 feature embeddings on CIFAR100

Fig. 7. **t-SNE visualizations of DRL embeddings on CIFAR100 under 'B0 Inc5' incremental setting.**

inference, the total parameter footprint scales minimally to $(1 + 0.006t)$B, where t represents the number of training stages, ensuring efficiency in both training and deployment phases. In terms of specific parameter counts, the pre-trained model (PTM) contains ∼102M parameters. For comparison, EASE introduces ∼0.29M learnable parameters, while our IPA network requires ∼0.62M learnable parameters. Notably, both methods maintain learnable parameters below 1% of the base model size, ensuring lightweight adaptation with minimal storage overhead. A key computational advantage of our approach lies in the inference process. While EASE requires executing inference through all previously trained models (T times) to obtain the final feature representation $\mathbf{F}_T$, our DRL approach performs inference only once through the final integrated model. This single-pass inference strategy substantially reduces computational overhead, memory requirements, and latency compared to methods requiring multiple forward passes. The combination of minimal parameter growth and efficient single-pass inference makes DRL particularly suitable for resource-constrained deployment scenarios where both storage and computational resources are limited.

Visualization. To qualitatively assess representation quality, we leverage t-SNE [76] to visualize the learned decision boundaries on the VTAB dataset across incremental stages. Fig. 6 illustrates the feature distributions on the VTAB dataset under the B0 Inc10 setting, comparing configurations with and without DAS. The visualizations reveal that DRL with DAS produces more compact and well-separated class clusters compared to the variant using conventional cross-entropy loss. Specifically, at stage 2 (Fig. 6d), the DAS-enhanced features maintain clear separation boundaries between classes from both incremental stages, demonstrating effective mitigation of catastrophic forgetting while accommodating new concepts. Additional visualizations on CIFAR100 (Fig. 7) under the B0 Inc5 setting further confirm these observations. The t-SNE plots show that DRL effectively distinguishes instances into their respective classes with minimal inter-class overlap. The superior clustering quality visually corroborates the quantitative performance advantages observed in our benchmark evaluations, providing intuitive evidence that DAS promotes more discriminative feature learning after the incremental learning process. These visualizations collectively demonstrate that our method successfully addresses the fundamental challenge of maintaining representation quality while sequentially incorporating new knowledge.

V. LIMITATIONS AND FUTURE WORK

While DRL demonstrates strong performance across multiple benchmarks, several limitations present opportunities for future research. The current transfer gate design, though effective, could benefit from more sophisticated feature fusion mechanisms such as attention-based dynamic weighting or hierarchical gating structures that adapt to class complexity. The virtual anchor parameter k is currently fixed throughout training; developing adaptive sampling strategies that adjust k based on class complexity or data distribution could further enhance performance. Additionally, while our experiments focus on medium-scale models, extending validation to larger architectures (e.g., ViT-L, ViT-H) would strengthen the evidence for scalability. Future work will also explore biological plausibility of the distillation mechanism and investigate applications in more challenging continual learning scenarios including task-incremental and domain-incremental settings. We will release code to facilitate reproduction and adoption by the research community.

VI. CONCLUSION

In this paper, we proposed a novel class-incremental learning (CIL) method, Discriminative Representation Learning (DRL), which consists of an Incremental Parallel Adapter (IPA) network and Decoupled Anchor Supervision (DAS). The IPA network achieves high efficiency and smooth representation shift through a lightweight adapter and a learnable transfer gate, enabling gradual model expansion while maintaining parameter efficiency. DAS enhances discriminative representation learning by decoupling constraints for positive and negative samples via a fixed anchor, which helps align feature spaces across incremental stages and mitigates optimization-inference inconsistency. Extensive experiments on six benchmarks demonstrate that DRL achieves state-of-the-art performance while maintaining high parameter efficiency.

REFERENCES

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

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[3] H. Ran, J. Liu, and C. Wang, "Surface representation for point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 942–18 952.

[4] J. Zhan, J. Liu, W. Tang, G. Jiang, X. Wang, B.-B. Gao, T. Zhang, W. Wu, W. Zhang, C. Wang, and Y. Xie, "Global meets local: Effective multi-label image classification via category-aware weak supervision," in *ACM International Conference on Multimedia*, ser. MM '22, vol. 30, Oct. 2022, p. 6318–6326.

[5] W. Li, J. Zhan, J. Wang, B. Xia, B.-B. Gao, J. Liu, C. Wang, and F. Zheng, "Towards continual adaptation in industrial anomaly detection," in *ACM International Conference on Multimedia*, 2022, pp. 2871–2880.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[8] J. Ning, L. Xie, C. Wang, Y. Bu, F. Xu, D.-W. Zhou, S. Lu, and B. Ye, "Rf-badge: Vital sign-based authentication via rfid tag array on badges," *IEEE Transactions on Mobile Computing*, vol. 22, no. 2, pp. 1170–1184, 2021.

[9] X. Tian, Z. Zhang, X. Tan, J. Liu, C. Wang, Y. Qu, G. Jiang, and Y. Xie, "Instance and category supervision are alternate learners for continual learning," in *International Conference on Computer Vision*, 2023, pp. 5596–5605.

[10] Z. Zhao, Z. Zhang, X. Tan, J. Liu, Y. Qu, Y. Xie, and L. Ma, "Rethinking gradient projection continual learning: Stability/plasticity feature space decoupling," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3718–3727.

[11] H. Zhang, B.-B. Gao, Y. Zeng, X. Tian, X. Tan, Z. Zhang, Y. Qu, J. Liu, and Y. Xie, "Learning task-aware language-image representation for class-incremental object detection," in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7096–7104.

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

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

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

[15] J. Dong, L. Wang, Z. Fang, G. Sun, S. Xu, X. Wang, and Q. Zhu, "Federated class-incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 164–10 173.

[16] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[17] M. A. P. Chamikara, P. Bertók, D. Liu, S. Camtepe, and I. Khalil, "Efficient data perturbation for privacy preserving and accurate data stream mining," *Pervasive and Mobile Computing*, vol. 48, pp. 1–19, 2018.

[18] S. T. Grossberg, *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*. Springer Science & Business Media, 2012, vol. 70.

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

[20] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3366–3375.

[21] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *The International Conference on Machine Learning*, 2018, pp. 4528–4537.

[22] A. Douillard, A. Ramé, G. Couairon, and M. Cord, "Dytox: Transformers for continual learning with dynamic token expansion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9285–9295.

[23] S. Yan, J. Xie, and X. He, "Der: Dynamically expandable representation for class incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3014–3023.

[24] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[25] F.-Y. Wang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Foster: Feature boosting and compression for class-incremental learning," in *European Conference on Computer Vision*, 2022, pp. 398–414.

[26] D.-W. Zhou, H.-L. Sun, J. Ning, H.-J. Ye, and D.-C. Zhan, "Continual learning with pre-trained models: a survey," in *International Joint Conference on Artificial Intelligence*, ser. IJCAI '24, 2024.

[27] J. Zheng, S. Qiu, and Q. Ma, "Learn or recall? revisiting incremental learning with pre-trained language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 14 848–14 877.

[28] W. Li, B.-B. Gao, B. Xia, J. Wang, J. Liu, Y. Liu, C. Wang, and F. Zheng, "Cross-modal alternating learning with task-aware representations for continual learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 5911–5924, 2023.

[29] D.-W. Zhou, H.-L. Sun, H.-J. Ye, and D.-C. Zhan, "Expandable subspace ensemble for pre-trained model-based class-incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 554–23 564.

[30] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[31] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio, "Gradient based sample selection for online continual learning." in *Advances in Neural Information Processing Systems*, 2019, pp. 11 816–11 825.

[32] Y. Liu, Y. Su, A.-A. Liu, B. Schiele, and Q. Sun, "Mnemonics training: Multi-class incremental learning without forgetting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 245–12 254.

[33] H. Zhao, H. Wang, Y. Fu, F. Wu, and X. Li, "Memory-efficient class-incremental learning for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5966–5977, 2021.

[34] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-gem," in *International Conference on Learning Representations*, 2018.

[35] H. Lin, S. Feng, X. Li, W. Li, and Y. Ye, "Anchor assisted experience replay for online class-incremental learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2217–2232, 2023.

[36] C. Simon, P. Koniusz, and M. Harandi, "On learning the geodesic path for incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1591–1600.

[37] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[38] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *European Conference on Computer Vision*, 2018, pp. 139–154.

[39] B. Zhao, X. Xiao, G. Gan, B. Zhang, and S.-T. Xia, "Maintaining discrimination and fairness in class incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 208–13 217.

[40] L. Yu, B. Twardowski, X. Liu, L. Herranz, K. Wang, Y. Cheng, S. Jui, and J. v. d. Weijer, "Semantic drift compensation for class-incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6982–6991.

[41] Y. Shi, K. Zhou, J. Liang, Z. Jiang, J. Feng, P. H. Torr, S. Bai, and V. Y. Tan, "Mimicking the oracle: An initial phase decorrelation approach for class incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 722–16 731.

[42] S. Dong, X. Gao, Y. He, Z. Zhou, A. C. Kot, and Y. Gong, "Ceat: Continual expansion and absorption transformer for non-exemplar class-incremental learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 4, pp. 3146–3159, 2025.

[43] D. Cheng, Y. Zhao, N. Wang, G. Li, D. Zhang, and X. Gao, "Efficient statistical sampling adaptation for exemplar-free class incremental learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11 451–11 463, 2024.

[44] K. Song, G. Liang, Z. Chen, and Y. Zhang, "Non-exemplar class-incremental learning by random auxiliary classes augmentation and mixed features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 7830–7843, 2024.

[45] H. Qu, H. Rahmani, L. Xu, B. Williams, and J. Liu, "Recent advances of continual learning in computer vision: An overview," *IET Computer Vision*, vol. 19, no. 1, p. e70013, 2025.

[46] X. Chen and X. Chang, "Dynamic residual classifier for class incremental learning," in *International Conference on Computer Vision*, 2023, pp. 18 743–18 752.

[47] B. Ni, X. Nie, C. Zhang, S. Xu, X. Zhang, G. Meng, and S. Xiang, "Moboo: Memory-boosted vision transformer for class-incremental learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11 169–11 183, 2024.

[48] L. Wang, J. Xie, X. Zhang, M. Huang, H. Su, and J. Zhu, "Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[49] M. D. McDonnell, D. Gong, E. Abbasnejad, and A. van den Hengel, "Premonition: Using generative models to preempt future data changes in continual learning," *arXiv preprint arXiv:2403.07356*, 2024.

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

[51] Y. Wang, Z. Huang, and X. Hong, "S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 5682–5695.

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

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

[54] D.-W. Zhou, Z.-W. Cai, H.-J. Ye, D.-C. Zhan, and Z. Liu, "Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need," *International Journal of Computer Vision*, pp. 1–21, 2024.

[55] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *The International Conference on Machine Learning*, 2019, pp. 2790–2799.

[56] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[57] Q. Li, Y. Peng, and J. Zhou, "Fcs: Feature calibration and separation for non-exemplar class incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 28 495–28 504.

[58] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *stat*, vol. 1050, p. 9, 2015.

[59] M. D. McDonnell, D. Gong, A. Parvaneh, E. Abbasnejad, and A. Van den Hengel, "Ranpac: Random projections and pre-trained models for continual learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 12 022–12 053.

[60] L. Zhao, X. Zhang, K. Yan, S. Ding, and W. Huang, "Safe: Slow and fast parameter-efficient tuning for continual learning with pre-trained models," in *Advances in Neural Information Processing Systems*, 2024.

[61] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[63] S. Chen, C. GE, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, 2022, pp. 16 664–16 678.

[64] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.

[65] D.-W. Zhou, Q.-W. Wang, H.-J. Ye, and D.-C. Zhan, "A model or 603 exemplars: Towards memory-efficient class-incremental learning," in *International Conference on Learning Representations*, 2023.

[66] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, L. Ma, S. Pu, and D.-C. Zhan, "Forward compatible few-shot class-incremental learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9046–9056.

[67] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

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

[69] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[70] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[71] Y. Zhang, Z. Yin, J. Shao, and Z. Liu, "Benchmarking omni-vision representation through the lens of visual realms," in *European Conference on Computer Vision*, 2022, pp. 594–611.

[72] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy *et al.*, "A large-scale study of representation learning with the visual task adaptation benchmark," *arXiv preprint arXiv:1910.04867*, 2019.

[73] D.-W. Zhou, Q.-W. Wang, Z.-H. Qi, H.-J. Ye, D.-C. Zhan, and Z. Liu, "Deep class-incremental learning: A survey," *arXiv preprint arXiv:2302.03648*, vol. 1, no. 2, p. 6, 2023.

[74] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[75] Y. Wang, Z. Ma, Z. Huang, Y. Wang, Z. Su, and X. Hong, "Isolation and impartial aggregation: A paradigm of incremental learning without interference," in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 10 209–10 217.

[76] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.