Geometry-driven OOD Detectors Are Class-Incremental Learners

Wangwang Jia[1,2]†, Zijian Gao[1,3]†, Tianjiao Wan[1,3], Yuan Cao[1,2], Yong Dou[1,2], Kele Xu[1,3]*
[1]College of Computer Science and Technology, National University of Defense Technology
[2]National Key Laboratory of Parallel and Distributed Computing, National University of Defense Technology
[3]State Key Laboratory of Complex & Critical Software Environment
{wangwangjia,gaozijian19,yuancao,yongdou,xukelele}@nudt.edu.cn

Abstract

Class-Incremental Learning (CIL) seeks to acquire new classes over time without erasing prior knowledge. While recent methods leverage pre-trained models (PTMs) to curb forgetting, they largely optimize the feature extractor and overlook the crucial classification head. In this work, we advance a simple view: if each task is equipped with a classifier that has the ability to both recognize in-distribution (IND) classes and reject out-of-distribution (OOD) inputs, CIL arises naturally—inputs are accepted only by heads that deem them in-distribution and rejected otherwise. Supported by rigorous theoretical and empirical studies, we find that this ability is characterized by ***Inter-class Separation*** *and* ***Intra-class Compactness****; lacking these, standard linear and cosine-similarity heads remain closed-set and fail to yield a usable OOD signal. To address this, we propose* ***GOD*** *(****G****eometry-driven* ***O****OD* ***D****etectors), which unifies IND recognition and OOD rejection in a single geometric space by replacing the learnable head with fixed Equiangular Tight Frame (ETF) anchors; an ETF loss enforces inter-class separation, and an ArcFace loss further tightens intra-class compactness. For efficiency, we further introduce a parameter-efficient hybrid architecture and an efficient inference strategy, thus reducing both parameter footprint and inference cost. Extensive experiments on multiple incremental settings and datasets show that GOD achieves state-of-the-art results.*[1]

1. Introduction

Conventional deep learning [9–11, 11, 21, 31, 32, 36–38, 43, 47, 52, 55, 80] operates in an offline batch setting, where the entire dataset is provided a priori. However, in open world settings, data often arrives in a stream. This necessitates a paradigm capable of incrementally acquiring knowledge of new classes, known as Class-Incremental Learning (CIL) [6, 15, 18, 40, 60, 62, 69]. The core challenge of CIL is catastrophic forgetting [16, 19, 20, 41, 42]: a sharp decline in performance on prior tasks after the model learns from new ones. To address this challenge, the CIL field is increasingly shifting its focus from traditional methods that train models from scratch [17, 33] towards leveraging large-scale pre-trained models (PTMs). PTMs, endowed with broad generalization from massive pre-training corpora, offer a strong substrate for continual adaptation and mitigating forgetting.

Due to the generalization of PTMs, existing works often freeze the pre-trained weights and adapt to incremental tasks using additional lightweight modules [18, 40, 54]. Along the classifier dimension, two designs are common. (i) Single expanding classifier [26, 57, 64, 67]: a single head grows as classes increase, but this typically causes representation forgetting and decision-level recency bias [4, 22, 35], shifting predictions toward new classes. (ii) Multi-fixed classifiers [1, 2, 25, 48]: task-specific heads define disjoint decision spaces and reduce recency bias, but they rely on a Task-ID predictor for routing, which itself forgets and becomes a major bottleneck.

Ideally, a classifier should not only recognize in-distribution (IND) classes for its own task but also reject unseen out-of-distribution (OOD) samples (i.e., provide a usable OOD signal). Standard softmax or cosine heads, optimized for closed-set IND classification, generally lack this property. If, instead, each task is equipped with a head that jointly produces class scores and an OOD acceptance decision, then test-time routing no longer depends on a fragile Task-ID predictor: the input is simply accepted by the head that deems it in-distribution and rejected otherwise. Under this view, CIL emerges naturally—the system grows by adding a new head for each new task, while previously trained heads maintain their own decision regions. This preserves disjoint decision spaces and avoids expanding a single shared classifier, thereby mitigating both representation forgetting and decision-level recency bias.

*Corresponding author. † Equal Contribution.
[1]Link to code-https://github.com/Wangwang-Jia/GOD.

In this paper, we introduce **G**eometry-driven **O**OD **D**etectors (**GOD**), a PTM-based, exemplar-free CIL (EFCIL) method that enables incremental learning by providing reliable OOD rejection. Our GOD handles classification and uncertainty estimation in a single geometric space, so that routing across tasks emerges from OOD decisions rather than a fragile Task-ID predictor. GOD replaces the learnable classifier with a fixed set of Equiangular Tight Frame (ETF) vectors used as class anchors. Training shifts from updating classifier weights to aligning features with these anchors, yielding a stable, globally consistent metric space across tasks.

We make two geometric assumptions: (i) *Inter-class Separation*—ETF anchors impose large, uniform angular margins between classes; and (ii) *Intra-class Compactness*—features of the same class concentrate around their anchor. Under these assumptions, the feature–anchor distance becomes a unified signal: it supports in-distribution classification and serves as a reliable uncertainty score for OOD rejection, outperforming existing classifiers (see Fig. 1(c)). To enforce these properties, we employ an ETF loss to realize inter-class separation, while an ArcFace [7] loss further tightens intra-class clusters and enlarges inter-class margins, improving separability around the ETF anchors.

However, multi-classifier CIL can cause parameters and inference cost to grow linearly as tasks accumulate. To address this, we adopt parameter-efficient Low-Rank Adaptation (LoRA) [27] and make an empirical observation: across tasks, LoRA updates in shallow blocks tend to converge to similar subspaces, whereas task-specific divergence concentrates in deeper blocks (see Fig. 3). Motivated by this, we factorize the adapters into a shared shallow LoRA stack and a task-specific deep stack. The shared stack is learned once and reused across tasks, while only small task-specific stacks are added per task, preserving adaptation capacity where variation resides and greatly reducing parameter overhead.

For inference, we further avoid activating all task heads. A universal EMA LoRA acts as a momentum-stabilized shared adapter: with a single forward pass, we obtain shared features and compute confidences for all heads (coarse mode). In a refined mode, we first identify the tasks corresponding to the top-k predicted classes and then re-encode only with their task-specific LoRA blocks, providing a tunable speed–accuracy trade-off.

In summary, the main contributions include:

- An analysis of existing classifiers, identifying key flaws: standard linear and cosine-similarity heads are closed-set and lack reliable OOD rejection ability.
- Supported by theoretical proofs, we propose GOD, which constructs a stable global metric via fixed ETF anchors, enabling unified IND classification and OOD rejection.
- A shared–specific LoRA decomposition for parameter-efficient tuning, together with a universal EMA–based two-mode router (coarse/refined) for inference efficiency.
- Extensive experiments showing that GOD consistently outperforms state-of-the-art PTM-based CIL methods on challenging benchmarks and settings.

2. Related Work

2.1. Pre-Trained Model-Based Class-Incremental Learning

Class-Incremental Learning (CIL) [3, 28, 68] aims to enable models to continually learn new knowledge while mitigating catastrophic forgetting of previously learned classes. Leveraging the transferable features provided by PTMs has become a prominent paradigm for tackling the CIL problem [60, 62, 73, 76]. Current PTM-based methods can be primarily categorized into three types: prompt-based [60, 62] adapt to new tasks by learning a set of task-specific prompts; representation-based [75, 76] focus on managing or adjusting representations in the feature space; and model-mixture-based [73] adapt by dynamically expanding their architecture. However, these mainstream approaches primarily focus on optimizing the feature extractor, while largely overlooking the design of the classifier and its critical role in the incremental learning process. In contrast, we focus on classifier design and propose GOD, a classifier that unifies IND classification with OOD uncertainty estimation.

2.2. Neural Collapse

Neural Collapse [45] is a phenomenon observed during terminal training phases, where last-layer features and classifier weights converge into a highly symmetric ETF structure. This theory provides an interpretable characterization of representation learning, revealing implicit geometric biases within deep models. Building on these insights, NC-inspired geometric regularization has informed transfer and generalization analyses [12, 13], architecture design [5], and incremental settings [13]. However, bringing these geometric properties to bear in strict EFCIL remains challenging: prior efforts [63] do not jointly achieve robust OOD uncertainty estimation and effective forgetting mitigation without an oracle Task-ID. This gap motivates a framework that constructs and preserves a stable, global metric space that is explicitly OOD-aware across sequential tasks.

3. Background and Notation

3.1. Problem Definition

In CIL, a model learns from a sequence of T incremental tasks. The training dataset for task t ($t = 1, \ldots, T$) is $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{N_t}$, comprising N_t samples. Each label y_i belongs to the task-specific class set $\mathcal{C}_t$ (with $|\mathcal{C}_t| = C_t$). The

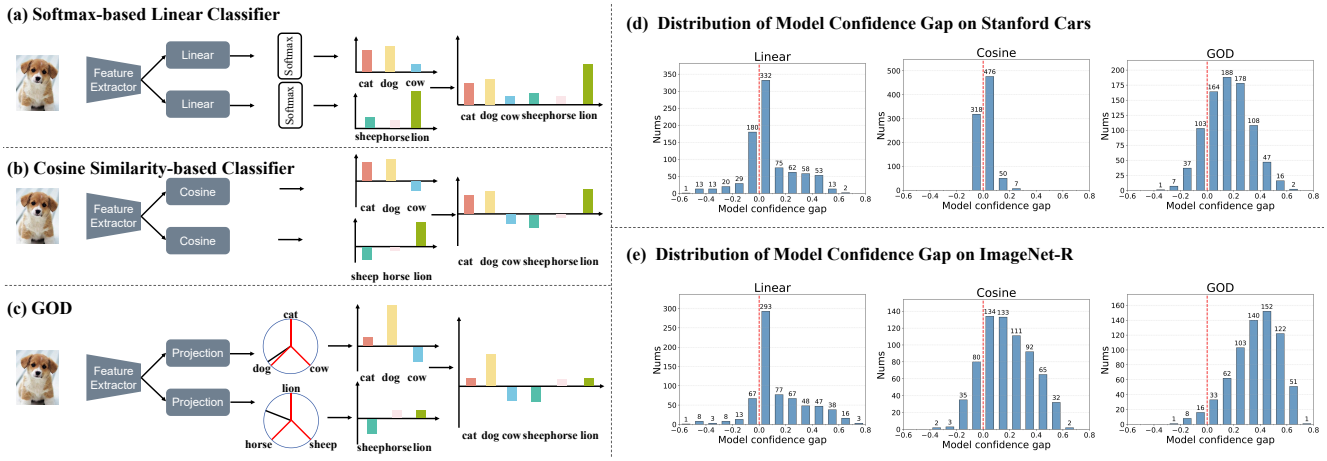


Figure 1. Comparison of inter-task score calibration for Linear, Cosine, and GOD classifiers. **(a-c)** Qualitative illustration of classifier logits for in-distribution (IND) and out-of-distribution (OOD) inputs. **(d-e)** Frequency distribution of the model confidence gap for the three classifiers, evaluated on two sequential tasks of Stanford Cars and ImageNet-R datasets.

class sets of different tasks are disjoint ($\forall t \neq t', \mathcal{C}_t \cap \mathcal{C}_{t'} = \emptyset$). Following PTM-based CIL [18, 40, 60, 62], we utilize a feature extractor $f(\cdot)$ initialized from a PTM and keep it frozen as a shared backbone. We adopt a replay-free setting [50, 77–79], where no instances from previous tasks are retained. The objective is to learn a unified classifier for all seen classes $\cup_{i=1}^{T} \mathcal{C}_i$ without accessing old data.

3.2. Low-Rank Adaptation

To bridge the domain gap between the generalized PTM representations and the downstream tasks, we use a lightweight, parameter-efficient fine-tuning method: Low-Rank Adaptation (LoRA) [27]. The Pre-Trained backbone $f(\cdot)$ consists of L transformer blocks, each containing a self-attention module and a Multi-Layer Perceptron (MLP) layer. Following recent works [39, 65], we apply LoRA only to the self-attention module within each block, leaving the MLP modules entirely frozen. This provides a complete set of adapter pairs for each task t, denoted as $\mathcal{L}_t = \{(A_{t,l}, B_{t,l})\}_{l=1}^{L}$, where l is the block index. The resulting task-specific model, $f_t(\cdot)$, is thus defined as the combination of the frozen backbone $f(\cdot)$ and its new, trainable adapters $\mathcal{L}_t$. LoRA aims to keep the original high-dimensional weights W (e.g., W_Q, W_K, W_V in the self-attention modules) frozen and inject a parallel, trainable path that models their task-specific updates ΔW_t as a low-rank decomposition. The forward pass of an adapted weight matrix for an input x is defined as:

$$h_t(x) = Wx + \Delta W_t x = Wx + (B_t A_t)x \tag{1}$$

where $A_t \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B_t \in \mathbb{R}^{d_{\text{out}} \times r}$ are the trainable low-rank matrices for task t, with a rank $r \ll \min(d_{\text{in}}, d_{\text{out}})$. Only the new matrices $\{A_t, B_t\}$ are trained for each task, while the entire original PTM backbone remains frozen.

4. Preliminary Analysis

4.1. Theoretical Foundation for OOD-Aware Classification

This section derives two core hypotheses for both IND classification and OOD rejection. Our analysis focuses on the geometry of the final feature extractor $z_t(\cdot) : \mathcal{X} \to \mathbb{R}^d$ (i.e., the last representation layer), seeking principles optimizable using only task t's data. We model the $k_{\text{IND}} = C_t$ IND classes (from task t) and $k_{\text{OOD}} = \sum_{m \neq t} C_m$ OOD classes (from other tasks). Their prototypes (mean feature vectors) are computed via the current model $z_t(\cdot)$:

$$\boldsymbol{\mu}_{\text{IND},c} = \mathbb{E}_{x \sim \mathcal{D}_{t,c}}[z_t(x)], \quad \forall c \in \mathcal{C}_t \tag{2}$$

$$\boldsymbol{\mu}_{\text{OOD},c'} = \mathbb{E}_{x \sim \mathcal{D}_{m,c'}}[z_t(x)], \quad \text{where } c' \in \mathcal{C}_m, m \neq t \tag{3}$$

Following prior works [44, 58], we first define a scoring function $ES_t(x)$ that is proportional to the data density:

$$ES_t(x) = \sum_{c \in \mathcal{C}_t} \exp\left(-\frac{1}{2} d_{M,t}(z_t(x), \boldsymbol{\mu}_{\text{IND},c})^2\right), \tag{4}$$

where $d_{M,t}(z, \boldsymbol{\mu}) = \sqrt{(z-\boldsymbol{\mu})^T \Sigma_t^{-1} (z-\boldsymbol{\mu})}$.

Here, $d_{M,t}(\cdot,\cdot)$ is the Mahalanobis distance using Σ_t, the shared covariance matrix of the IND classes. We then define D_t, a measure of the classifier's inherent ability to separate IND from OOD, as the expectation difference in $ES_t(x)$ between IND and OOD samples:

$$D_t = \mathbb{E}_{x \sim \mathcal{D}_t}[ES_t(x)] - \mathbb{E}_{x \sim \bigcup_{m \neq t} \mathcal{D}_m}[ES_t(x)] \tag{5}$$

However, it is unoptimizable in our replay-free setting as it requires inaccessible $\boldsymbol{\mu}_{\text{OOD},c'}$. To find a computable optimization proxy, we derive an upper bound on D_t. By applying the triangle inequality to the OOD-to-IND distances, we

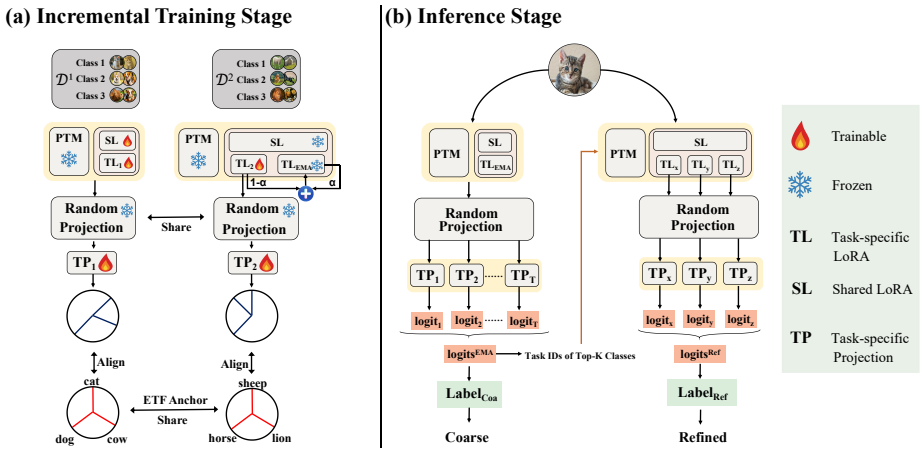


Figure 2. The GOD framework. **(a) Incremental Training:** SL and the PTM are trained on the first task and then frozen. For each subsequent task, we keep the PTM and SL fixed and only train a new task-specific TL_t and projection head TP_t, while TL_{ema} is updated via EMA. **(b) Inference:** In the Coarse mode, SL and TL_{ema} produce logits for all task heads in a single pass. In the Refined mode, TL_{ema} selects the Top-k task-specific modules (e.g., TL_x, TL_y, TL_z) for re-encoding, enabling efficient prediction via sparse activation.

relax the bound as follows (See the supplementary materials for Proof.):

Theorem 1. *Let $c_0(c')$ be the index of the closest IND prototype to any OOD prototype $\boldsymbol{\mu}_{OOD,c'}$, defined as:*

$$c_0(c') = \arg\min_{c \in \mathcal{C}_t} d_{M,t}(\boldsymbol{\mu}_{OOD,c'}, \boldsymbol{\mu}_{IND,c}). \tag{6}$$

The separation metric D_t is bounded as:

$$D_t \leq \frac{k_{IND}}{2k_{OOD}} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} d_{M,t}(\boldsymbol{\mu}_{OOD,c'}, \boldsymbol{\mu}_{IND,c_0(c')}) + \frac{1}{2} \sum_{c \in \mathcal{C}_t} d_{M,t}(\boldsymbol{\mu}_{IND,c_0(c')}, \boldsymbol{\mu}_{IND,c}). \tag{7}$$

Theorem 1 provides the blueprint. The bound (Eq. (7)) is uncomputable as it depends on unknown $\boldsymbol{\mu}_{\text{OOD},c'}$, but its structure reveals two geometric principles for raising it. First, the bound is raised by maximizing the IND inter-class distance (the second term). Second, minimizing the covariance Σ_t by enforcing compactness around anchors amplifies all Mahalanobis distances. This makes maximizing angular separation via ETF a highly effective proxy for the Mahalanobis-based objective. To summarize, the theoretical foundation leads to two core hypotheses:

Hypothesis 1 (Inter-class Separation): Maximizing the angular separation between class anchors increases the inter-class margins, tightening the generalization bound.

Hypothesis 2 (Intra-class Compactness): Minimizing the feature covariance within each class amplifies Mahalanobis distances to other classes, improving separability under distribution shift.

4.2. Limitations of Conventional Classifiers

These two core hypotheses define an ideal geometry for IND classification and OOD rejection—a geometry that traditional classifiers cannot construct. Softmax classifiers exhibit inherent OOD overconfidence (Fig. 1a), while Cosine classifiers learn only local decision boundaries, not the global anchors our theory demands. Our GOD classifier bridges this gap, instantiating the requisite geometric foundation via ETF architecture. Figure 1(d)(e) quantifies these failures. We trained Linear, Cosine, and GOD classifiers on two sequential tasks from Stanford Cars [29] and ImageNet-R [23], plotting the distribution of the confidence gap—maximum confidence from the correct task model minus that from the incorrect task model. Negative values indicate Task ID failures. Linear and Cosine classifiers exhibit numerous failures; even when correct, their confidence gaps are extremely low and clustered near zero, signaling inevitable degradation as more tasks are added. In contrast, GOD achieves far fewer failures and substantially larger gaps, showing its intrinsic calibration: high confidence for compact IND clusters and low confidence for OOD samples, satisfying both hypothesis 1 and 2.

5. GOD: Geometry-driven OOD Detectors

Supported by the theoretical foundation and empirical studies, we attribute the success of GOD to three key components: a geometry-driven training paradigm, a PTM-tailored parameter-efficient hybrid architecture, and an efficient inference strategy. Figure 2 illustrates the overall framework of GOD and the pseudo-algorithm is included in the supplementary material.

5.1. Geometry-driven Training Paradigm

We demonstrated that the effectiveness of a task-specific head in CIL fundamentally hinges on two geometric properties: *inter-class separation* and *intra-class compactness*. GOD operationalizes these design principles by first *fixing*

an ideal target geometry and then *learning* to align features to this geometry, instead of learning both the feature space and classifier weights jointly. For each task t, we construct a set of normalized ETF anchors

$$E_t = \{e_{t,1}, \dots, e_{t,C_t}\}, \tag{8}$$

which satisfy

$$\langle e_{t,i}, e_{t,j} \rangle = \begin{cases} 1, & i = j, \\ -\frac{1}{C_t - 1}, & i \neq j, \end{cases} \tag{9}$$

thus forming a maximally separated, symmetric configuration on the unit hypersphere. These anchors play the role of "ideal prototypes" for each class, encoding the desired large and uniform angular margins that Section 4 identified as crucial for reliable OOD rejection.

Given the task-specific feature extractor $f_t(\cdot)$ using the corresponding low-rank adapter, GOD uses a two-stage linear projection followed by ℓ_2 normalization to obtain the final representation $z_t(x) \in \mathbb{S}^{d-1}$:

$$z_t(x) = \frac{\mathrm{TP}_t(\mathrm{RP}(f_t(x)))}{\|\mathrm{TP}_t(\mathrm{RP}(f_t(x)))\|}. \tag{10}$$

Here, a shared, frozen Random Projection (RP) first lifts the features into a higher-dimensional space, enhancing linear separability across tasks. The task-specific projection head TP_t then learns to align the projected features with the ETF anchors. Both RP and TP_t are implemented as single linear layers, making the overall design simple and scalable.

The logit of class $c \in \{1, \dots, C_t\}$ is defined as the cosine similarity between $z_t(x)$ and $e_{t,c}$:

$$s_{t,c}(x) = \langle z_t(x), e_{t,c} \rangle, \tag{11}$$

and we denote the task-specific logit vector as

$$s_t(x) = (s_{t,1}(x), \dots, s_{t,C_t}(x))^\top. \tag{12}$$

At test time, all task heads are combined into a single global logit vector $s(x) = \mathrm{concat}_{t=1}^{T}(s_t(x))$, which realizes the multi-head OOD detector.

By hard-coding the classifier as an ETF and using a nearest-anchor decision rule, GOD inherently satisfies three of the four Neural Collapse (NC) [45] conditions: (NC2) ETF geometry, (NC3) self-duality, and (NC4) nearest class-center decision boundaries. (See the supplementary materials for details.) Training thus only needs to enforce (NC1)—within-class variability collapse—by driving $z_t(x)$ to concentrate around the corresponding anchors. This reduces the learning problem to shaping the feature distribution under a *fixed* optimal geometry, directly aligning with our theoretical characterization of IND/OOD behavior in Section 4. To explicitly instantiate Hypothesis 1 (inter-class separation) and Hypothesis 2 (intra-class compactness), we use a composite objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{etf}} + \mathcal{L}_{\text{arc}}. \tag{13}$$

Inter-class separation. We first employ a temperature-scaled cross-entropy loss

$$\mathcal{L}_{\text{etf}}(x_i, y_i) = -\log \frac{\exp(s_{t,y_i}(x_i)/\tau)}{\sum_{j=1}^{C_t} \exp(s_{t,j}(x_i)/\tau)}, \tag{14}$$

where $s_{t,y_i}(x_i)$ is the logit of the ground-truth class y_i and τ is a temperature parameter. This term encourages features to align with the ETF anchors and fully exploits the global angular margins encoded by the ETF, thus reinforcing inter-class separation.

Intra-class compactness. To further tighten the clusters around each anchor, we adopt the ArcFace loss $\mathcal{L}_{\text{arc}}$ [7]:

$$\mathcal{L}_{arc}(x_i, y_i) = -\log \frac{e^{s \cdot \cos(\theta_{y_i} + m)}}{e^{s \cdot \cos(\theta_{y_i} + m)} + \sum_{j=1, j \neq y_i}^{C_t} e^{s \cdot s_{t,j}(x_i)}}, \tag{15}$$

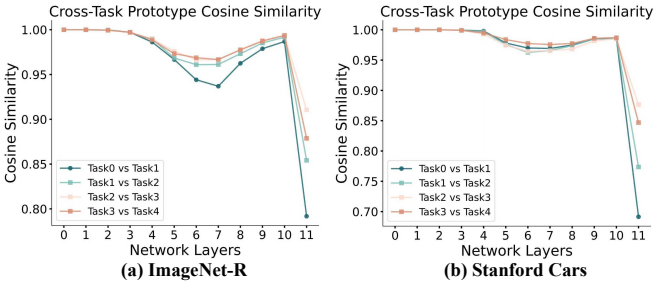
where θ_{y_i} is the angle between $z_t(x_i)$ and e_{t,y_i}, s is a scaling factor, and m is an additive angular margin. By explicitly enlarging the angular margin around each anchor, this term directly enforces intra-class compactness and effectively boosts local inter-class separation, thereby jointly supporting Hypotheses 1 and 2.

In summary, GOD realizes a geometry-aware training paradigm that is theoretically grounded in Section 4 and directly aligned with the high-level goal set out in Section 1: constructing task heads that are intrinsically capable of acting as reliable IND/OOD detectors, which we formally prove in the supplementary materials provides a deterministic guarantee for perfect IND classification and OOD rejection.

5.2. Parameter-efficient Hybrid Architecture

While assigning a full LoRA stack to every task provides strong adaptation ability, the parameter count grows linearly with the number of tasks. To reconcile scalability with performance, in Figure 3, we revisit the representational analysis and examine the cosine similarity of task-specific prototypes across depth. Empirically, we observe that prototypes in shallow transformer blocks exhibit high cross-task similarity, suggesting that LoRA updates there largely capture task-agnostic features. In contrast, prototypes in deeper blocks show much larger diversity, indicating that task-specific variations concentrate in deeper layers. This depth-wise heterogeneity naturally motivates a *hybrid* sharing scheme. Concretely, we partition the L transformer blocks at depth k:

- LoRA modules in blocks $l \leq k$ are shared across all tasks and form a single *Shared LoRA* (SL) module.



(a) ImageNet-R **(b) Stanford Cars**

Figure 3. Analysis of feature specificity across network layers on Stanford Cars (a) and Imagenet-R (b). The plots show the average cosine similarity of layer-wise prototypes for the same classes across models trained sequentially on five tasks.

- LoRA modules in blocks $l > k$ remain task-specific and form the *Task-specific LoRA* (TL) module.

The SL module is trained only on the first task ($t = 1$) and then frozen. For each subsequent task $t > 1$, we instantiate a lightweight TL_t module, initialized from TL_{t-1} to provide a strong warm start. This architecture concentrates task-specific capacity where it is most needed (deep layers), while the shared shallow layers capture the generic representations that are identified as transferable across tasks.

5.3. Efficient Inference Strategy

Activating all TL_t modules during inference would lead to $O(T)$ computational cost, which is undesirable in long task sequences. To maintain efficiency, we introduce a universal EMA-based adapter, denoted as $\mathrm{TL}^{\mathrm{ema}}$, which aggregates information from all task-specific TL_t modules into a single global router. $\mathrm{TL}^{\mathrm{ema}}$ is initialized with TL_1's parameters Θ_1 and updated after each epoch for $t \geq 2$ via

$$\Theta^{\mathrm{ema}} \leftarrow \alpha \cdot \Theta^{\mathrm{ema}} + (1 - \alpha) \cdot \Theta_t, \tag{16}$$

where Θ_t denotes the parameters of TL_t and α is the EMA momentum. $\mathrm{TL}^{\mathrm{ema}}$ accumulates a stable, low-variance approximation of task-specific knowledge. At inference time, we consider two modes that realize a controllable accuracy–efficiency trade-off.

Coarse mode. We process an input x with the frozen PTM augmented by SL and $\mathrm{TL}^{\mathrm{ema}}$, obtaining a shared representation $f^{\mathrm{ema}}(x)$. This representation is then passed through the shared RP and all task-specific heads TP_t to yield the logit vector $\mathrm{logits}^{\mathrm{EMA}}(x)$. The resulting prediction, denoted as $\mathrm{Label}_{\mathrm{Coa}}$, requires only one forward pass and already exhibits strong IND/OOD behavior.

Refined mode. For scenarios where accuracy is paramount, we refine the prediction using sparse task activation. Specifically, we first use $\mathrm{logits}^{\mathrm{EMA}}(x)$ to select the top-k classes and their associated task IDs, forming a small candidate set $\mathcal{T}_{\text{top-}k}$. We then reuse the intermediate features after SL, re-encode them with the corresponding TL_t modules for $t \in \mathcal{T}_{\text{top-}k}$, and pass the resulting features through RP and TP_t to obtain refined logits $\mathrm{logits}^{\mathrm{Ref}}(x)$ and the final prediction $\mathrm{Label}_{\mathrm{Ref}}$.

6. Experiment

6.1. Experiment setting

Datasets. To evaluate the performance of our method, we conduct experiments on four commonly used benchmarks: CIFAR-100 [30], ImageNet-A [24], ImageNet-R [23], and Stanford Cars [29]. All classes are arranged in a fixed order and the model is trained on an equal number of classes in both the first task and each incremental task. For all datasets, we consider incremental settings with $T = 10$ and $T = 5$ tasks. More results on 20-task setting are in the supplementary file.

Implementation details. All methods are implemented in PyTorch [46] using the PILOT toolbox [53], and trained on an NVIDIA RTX 4090 GPU. Consistent with recent PTM-based CIL works [61, 71], we adopt ViT-B/16 pre-trained on ImageNet-21K [8] as the backbone, which contains 12 transformer blocks in total. Following [49], we fix the random seed to 1993 and all compared methods use the same sequence for fair comparison. As for our GOD, the first 9 transformer blocks are frozen during incremental stages and constitute the SL module, while the last 3 blocks are adapted by task-specific TL modules. For Refined inference, the Top-k hyperparameter is set to $k = 3$ by default. Additional implementation details are provided in the supplementary material.

Comparison methods. To establish the effectiveness of GOD, we compare it against state-of-the-art (SOTA) PTM-based CIL methods, including DualPrompt [59], L2P [61], CODA-Prompt [51], LAE [14], DS-AL [81], SimpleCIL [75], Aper [71], EASE [72], LORA-DRS [39], SD-LoRA [65], and NC-CIPM [63]. For Finetune, LwF [34], and Aper [71], we construct a fair comparison by using a frozen PTM with a trainable LoRA module as the backbone. Furthermore, we also benchmark our approach against typical replay-based CIL methods, such as iCaRL [49], DER [66], FOSTER [56], and MEMO [70] in the supplementary file.

Evaluation Metrics. Following the widely used benchmarks [53, 74], two metrics are adopted for evaluation. The overall performance is evaluated by the average incremental accuracy $\bar{A}(\%) : \bar{A} = \frac{1}{T}\sum_{t=1}^{T} A_t$ where T and A_t respectively denote the number of incremental tasks and the test accuracy on all seen classes at task t. Another metric is the last-task accuracy $A_T(\%)$, which denotes the final average accuracy on all classes.

6.2. State-of-the-art Performance Comparison

We first present a comprehensive comparison with SOTA PTM-based CIL methods in Table 1. Overall, GOD con-

Table 1. Comparison of average incremental accuracy $\bar{A}(\%)$, last-task accuracy $A_T(\%)$, and mean performance of EFCIL methods across different numbers of tasks T and datasets. The best results among existing methods are highlighted in blue.

Metric	Methods	CIFAR-100		ImageNet-A		ImageNet-R		Stanford Cars	
		$T=10$	5	$T=10$	5	$T=10$	5	$T=10$	5
$\bar{A}$	Finetune	87.52	90.51	50.45	57.23	74.94	78.15	60.93	68.098
	LwF (TPAMI 2018) [34]	82.88	88.1	56.25	59.867	72.355	80.812	58.19	66.01
	DualPrompt (ECCV 2022) [59]	88.86	89.78	57.08	59.51	74.56	74.93	59.32	63.424
	L2P (CVPR 2022) [61]	89.48	91.02	53.99	56.45	77.78	77.86	67.42	71.918
	CODA-Prompt (CVPR 2023) [51]	91.19	92.20	62.83	67.11	77.56	80.44	44.779	67.107
	LAE (ICCV 2023) [14]	86.97	88.50	55.66	58.07	74.20	72.91	54.972	64.67
	DS-AL (AAAI 2024) [81]	83.50	88.82	61.75	63.40	78.38	78.80	62.114	62.483
	Aper (IJCV 2024) [71]	90.91	91.56	59.48	61.49	76.55	78.91	49.591	49.63
	EASE (CVPR 2024) [72]	**92.01**	92.24	62.89	68.17	81.38	82.28	50.525	56.302
	SimpleCIL (IJCV 2024) [75]	82.31	81.12	59.33	58.09	67.09	65.09	49.368	47.55
	NC-CIPM (AAAI 2025) [63]	90.94	92.35	64.15	65.67	78.56	81.50	71.42	72.98
	LORA-DRS (CVPR 2025) [39]	91.49	**92.38**	**69.64**	**71.33**	81.76	82.56	63.22	68.91
	SD-LoRA (ICLR 2025) [65]	91.55	92.18	58.41	63.46	**82.88**	**83.68**	**73.05**	**79.59**
	GOD (All)	**92.33**	**92.34**	**70.52**	**73.36**	**85.18**	**86.74**	**77.32**	**85.00**
	GOD (Coarse)	**92.15**	**92.26**	**69.97**	**71.89**	**83.84**	**85.31**	**74.57**	**84.93**
	GOD (Refined)	**92.87** (+0.86)	**92.78** (+0.40)	**70.50** (+0.86)	**73.39** (+2.06)	**85.41** (+2.53)	**86.90** (+3.22)	**77.10** (+4.05)	**85.23** (+5.64)
A_T	Finetune	82.06	86.31	40.62	47.14	68.63	73.97	45.04	58.41
	LwF (TPAMI 2018) [34]	77.57	84.28	40.22	45.89	69.55	73.27	40.80	53.56
	DualPrompt (ECCV 2022) [59]	84.23	84.76	47.93	49.18	69.10	70.37	44.19	50.97
	L2P (CVPR 2022) [61]	84.47	86.27	45.49	48.52	72.25	73.73	53.97	63.25
	CODA-Prompt (CVPR 2023) [51]	87.24	88.67	51.02	56.35	73.10	76.40	32.34	57.04
	LAE (ICCV 2023) [14]	81.13	82.76	47.73	50.03	69.83	71.05	41.17	53.01
	DS-AL (AAAI 2024) [81]	86.05	85.91	52.67	51.02	77.48	76.55	12.73	11.19
	Aper (IJCV 2024) [71]	85.81	87.58	55.69	59.91	72.05	74.95	38.52	40.30
	EASE (CVPR 2024) [72]	**87.25**	**89.22**	51.74	57.93	76.00	78.05	37.76	44.78
	SimpleCIL (IJCV 2024) [75]	76.21	76.21	49.24	49.24	61.35	61.35	38.26	38.26
	NC-CIPM (AAAI 2025) [63]	87.24	88.41	54.71	57.47	73.63	76.78	61.82	64.6
	LORA-DRS (CVPR 2025) [39]	85.89	88.28	**58.76**	**62.28**	**77.58**	78.18	51.23	59.41
	SD-LoRA (ICLR 2025) [65]	87.09	88.81	48.39	54.25	77.08	**78.47**	**60.30**	**71.50**
	GOD (All)	**87.06**	**88.24**	**59.64**	**63.86**	**79.15**	**81.78**	**67.58**	**77.75**
	GOD (Coarse)	**87.28**	**87.58**	**57.67**	**59.78**	**79.08**	**80.13**	**65.98**	**76.49**
	GOD (Refined)	**88.11** (+0.86)	**89.50** (+0.28)	**59.38** (+0.62)	**63.99** (+1.71)	**79.20** (+1.62)	**81.59** (+3.12)	**66.8** (+6.50)	**77.87** (+6.37)

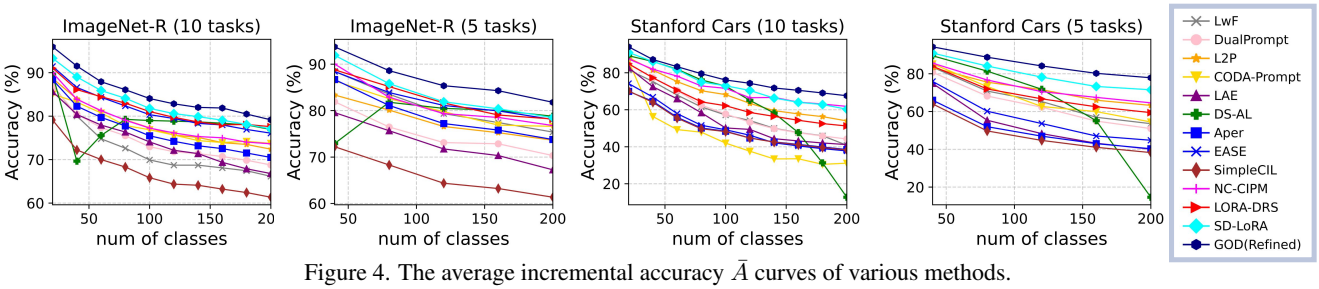


Figure 4. The average incremental accuracy $\bar{A}$ curves of various methods.

sistently achieves state-of-the-art performance. On relatively simple benchmarks such as CIFAR-100, our method is highly competitive (e.g., 92.87% $\bar{A}$), but the margin over strong baselines is small, which we attribute to the PTM's strong inherent generalization on this dataset. In contrast, GOD's advantage becomes clear on more complex, fine-grained benchmarks: on ImageNet-R (5 tasks), **GOD (Refined)** achieves 85.41% $\bar{A}$ (+2.53% over the runner-up), and on Stanford Cars (5 tasks) it reaches 85.23% $\bar{A}$ (+5.64%). These results highlight that our geometry-driven head yields substantial gains precisely when the PTM's generic features are insufficient.

Table 1 also illustrates the flexibility of our inference design. **GOD(Coarse)** delivers highly competitive performance with a single backbone forward pass, already outperforming many SOTA methods on ImageNet-R. The **GOD (Refined)** mode further applies Top-k sparse activation to achieve accuracy (e.g., 86.90% on ImageNet-R with 5 tasks) that is nearly identical to the expensive **GOD(All)** variant (86.74%), while substantially reducing computational cost, where **GOD (All)** denotes the variant that activates all task-specific LoRA blocks for each input.

Furthermore, Fig. 4 plots the evolution of average incremental accuracy $\bar{A}$ as the number of tasks increases. Using the efficient **GOD (Refined)** configuration, we observe that, on ImageNet-R and Stanford Cars, the performance gap between GOD and strong competitors (LORA-DRS, SD-LoRA, EASE) consistently widens as more tasks are learned. This trend indicates that GOD is markedly more effective over long task sequences. Additional results on other PTM backbones and comparisons with replay-based methods are provided in the supplementary material.

6.3. Computational Cost and Memory Comparison

Beyond accuracy, Table 2 compares the computational and memory footprint of GOD with strong PTM-based CIL baselines. Despite operating on the same ViT-B/16 backbone, GOD requires only 2.64M trainable parameters, reflecting the effect of the parameter-efficient hybrid architecture where generic shallow LoRA blocks are shared and only lightweight task-specific TL modules are added. While the total parameter count is similar across methods due to

Table 2. Computational cost and memory comparison on ImageNet-A ($T = 5$).

Metrics	DS-AL	Aper	EASE	NC-CIPM	LORA-DRS	SD-LoRA	GOD(Coarse)	GOD(Refined)
Trainable Params (#M)	4.73	4.73	4.73	4.73	4.73	4.91	2.64	2.64
All Params (#M)	95.74	176.33	94.10	91.83	95.63	172.08	96.33	100.09
Inference Time (s)	5.17	5.45	10.45	3.03	3.76	4.69	3.79	9.24

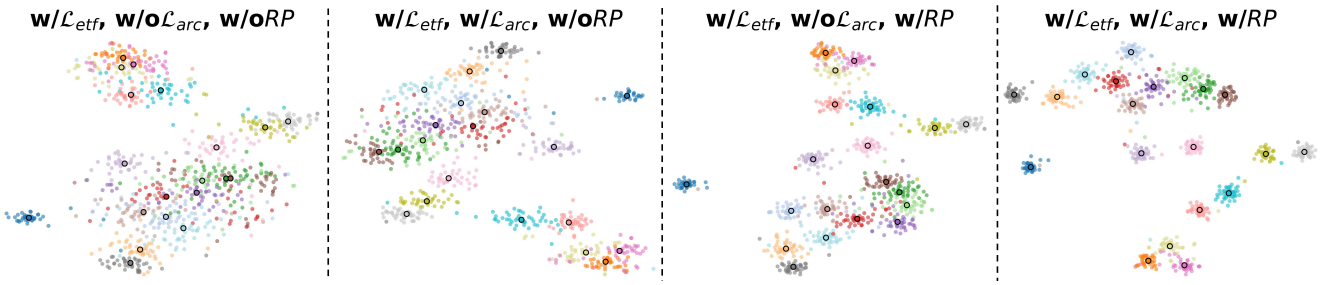


Figure 5. t-SNE visualizations of GOD variants on Stanford Cars after training on 20 classes in the first task.

Table 3. Average incremental accuracy $\bar{A}(\%)$ comparison of different values of Top-k across various settings.

k	Stanford Cars		Imagenet-R	
	$T = 10$	5	$T = 10$	5
2	76.65	85.24	84.85	86.65
3	77.10	**85.23**	**85.41**	86.90
4	77.17	85.07	85.14	86.77
5	**77.24**	85.04	85.16	**86.99**

the dominant PTM, the incremental overhead per task for GOD is markedly lower. In terms of inference time, GOD (Coarse) remains competitive with the most efficient baselines, whereas GOD (Refined) trades a moderate increase in latency for the accuracy gains reported in Table 1.

6.4. Geometry-driven Feature Visualization

To further understand how our geometry-driven design shapes the representation space, Fig. 5 visualizes the learned features for four variants of GOD (Refined). From left to right, we incrementally enable the three components in GOD: (i) $\mathcal{L}_{etf}$ *only*, where features are trained solely with $\mathcal{L}_{\text{etf}}$; (ii) $\mathcal{L}_{etf}$ + $\mathcal{L}_{arc}$, which additionally introduces the ArcFace loss $\mathcal{L}_{\text{arc}}$; (iii) $\mathcal{L}_{etf}$ + *RP*, which combines $\mathcal{L}_{\text{etf}}$ with the Random Projection module but without $\mathcal{L}_{\text{arc}}$; and (iv) the full $\mathcal{L}_{etf}$ + $\mathcal{L}_{arc}$ + *RP* configuration used in GOD.

With $\mathcal{L}_{\text{etf}}$ only (far left), class regions are roughly separated but remain diffuse, and different classes still exhibit noticeable overlap. Adding $\mathcal{L}_{\text{arc}}$ (second) tightens clusters around their ETF anchors, yet inter-class confusion remains, indicating that intra-class compactness alone is insufficient. In contrast, adding RP on top of $\mathcal{L}_{\text{etf}}$ (third) improves global separation, but clusters are still relatively loose. The full configuration (far right), which combines $\mathcal{L}_{\text{etf}}$, $\mathcal{L}_{\text{arc}}$, and RP, yields both compact and well-separated clusters. This progressive improvement aligns with Hypotheses 1 and 2: $\mathcal{L}_{\text{etf}}$ and RP primarily enhance inter-class separation, $\mathcal{L}_{\text{arc}}$ enforces intra-class compactness, and their combination realizes the "high intra-class cohesion, low inter-class coupling" regime where our theoretical analysis guarantees reliable IND classification and OOD-aware rejection across tasks. (See the supplementary material for more ablation analysis)

6.5. Sensitivity Analysis

We conducted a sensitivity analysis indicating that GOD is highly robust. Table 3 examines the Top-k hyperparameter in our sparse task-activation scheme: $\bar{A}$ changes very little when k varies from 2 to 5, with the best results around $k = 3 \sim 5$. We set $k = 3$ by default as a good efficiency–accuracy trade-off. Additional ablations on the random projection dimension d_B, ArcFace margin m, EMA momentum α, and the number of shared layers Num_{SL} are provided in the supplementary material, and consistently support the effectiveness and robustness of GOD.

7. Conclusion

Incremental learning is a key capability for intelligent systems in evolving environments. Existing PTM-based CIL methods, however, largely focus on improving in-distribution classification accuracy with closed-set heads, while overlooking the classifier's role as an out-of-distribution gate. In contrast, we explicitly design task heads to be OOD-aware classifiers, so that extending the system with new tasks does not require expanding a single shared head and preserves previously learned decision regions. We instantiate this view with GOD, a geometry-driven OOD detector that unifies IND recognition and OOD rejection in a single metric space by enforcing inter-class separation and intra-class compactness. Combined with a parameter-efficient hybrid LoRA architecture and a coarse–refined inference strategy, GOD achieves state-of-the-art performance on challenging benchmarks while keeping parameter and inference overhead low. These results underscore the importance of classifier geometry and suggest geometry-driven designs as a promising direction for future continual and open-world learning.

Acknowledgments

This work is supported by National Science and Technology Major Project (2023ZD0121101), National University of Defense Technology (ZZCX-ZZGC-01-04) and Major Fundamental Research Project of Hunan Province (2025JC0005).

References

[1] 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. 1

[2] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017. 1

[3] Arjun Ashok, KJ Joseph, and Vineeth N Balasubramanian. Class-incremental learning with cross-space clustering and controlled transfer. In *European conference on computer vision*, pages 105–122. Springer, 2022. 2

[4] Xusheng Cao, Haori Lu, Linlan Huang, Xialei Liu, and Ming-Ming Cheng. Generative multi-modal models are good class incremental learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28706–28717, 2024. 1

[5] Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of machine learning research*, 23(114):1–103, 2022. 2

[6] Yawen Cui, Jian Zhao, Zitong Yu, Rizhao Cai, Xun Wang, Lei Jin, Alex C. Kot, Li Liu, and Xuelong Li. Cmoa: Contrastive mixture of adapters for generalized few-shot continual learning. *IEEE Transactions on Multimedia*, 27:5533–5547, 2025. 1

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[9] Songcheng Du, Yang Zou, Jiaxin Li, Mingxuan Liu, Ying Li, Changjing Shang, and Qiang Shen. Pansharpening for thin-cloud contaminated remote sensing images: a unified framework and benchmark dataset. In *Proceedings of AAAI 2026*. 2026. 1

[10] Songcheng Du, Yang Zou, Jiaxin Li, Mingxuan Liu, Ying Li, Changjing Shang, and Qiang Shen. Pansharpening for thin-cloud contaminated remote sensing images: a unified framework and benchmark dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3696–3704, 2026.

[11] Songcheng Du, Yang Zou, Zixu Wang, Xingyuan Li, Ying Li, Changjing Shang, and Qiang Shen. Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. *International Journal of Computer Vision*, 2026. 1

[12] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021. 2

[13] Tomer Galanti, András György, and Marcus Hutter. Generalization bounds for transfer learning with pretrained classifiers. *CoRR*, 2022. 2

[14] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11483–11493, 2023. 6, 7

[15] Zijian Gao, Kele Xu, Huiping Zhuang, Li Liu, Xinjun Mao, Bo Ding, Dawei Feng, and Huaimin Wang. Less confidence, less forgetting: Learning with a humbler teacher in exemplar-free class-incremental learning. *Neural Networks*, 179:106513, 2024. 1

[16] Zijian Gao, Xingxing Zhang, Kele Xu, Xinjun Mao, and Huaimin Wang. Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks. In *Advances in Neural Information Processing Systems*, pages 128462–128488. Curran Associates, Inc., 2024. 1

[17] Zijian Gao, Shanhao Han, Xingxing Zhang, Kele Xu, Dulan Zhou, Xinjun Mao, Yong Dou, and Huaimin Wang. Maintaining fairness in logit-based knowledge distillation for class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16763–16771, 2025. 1

[18] Zijian Gao, Wangwang Jia, Xingxing Zhang, Dulan Zhou, Kele Xu, Feng Dawei, Yong Dou, Xinjun Mao, and Huaimin Wang. Knowledge memorization and rumination for pretrained model-based class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20523–20533, 2025. 1, 3

[19] Zijian Gao, Kele Xu, Xingxing Zhang, Huiping Zhuang, Tianjiao Wan, Bo Ding, Xinjun Mao, and Huaimin Wang. Rethinking obscured sub-optimality in analytic learning for exemplar-free class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025. 1

[20] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Computer Science*, 84 (12):1387–91, 2013. 1

[21] MH Guo, CZ Lu, ZN Liu, MM Cheng, and SM Hu. Visual attention network: Computational visual media. 2023. 1

[22] Jiangpeng He. Gradient reweighting: Towards imbalanced class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16668–16677, 2024. 1

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu,

Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. 4, 6

[24] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021. 6

[25] 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. 1

[26] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019. 1

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2, 3

[28] 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. 2

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 4, 6

[30] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009. 6

[31] Hesong Li and Ying Fu. Fcdfusion: A fast, low color deviation method for fusing visible and infrared image pairs. *Computational Visual Media*, 11(1):195–211, 2025. 1

[32] Hesong Li, Ziqi Wu, Ruiwen Shao, Tao Zhang, and Ying Fu. Noise calibration and spatial-frequency interactive network for stem image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Conference*, pages 21287–21296, 2025. 1

[33] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1

[34] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(12):2935–2947, 2017. 6, 7

[35] Guoqiang Liang, Zhaojie Chen, Zhaoqiang Chen, Shiyu Ji, and Yanning Zhang. New insights on relieving task-recency bias for online class incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5): 3451–3464, 2023. 1

[36] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9552–9566, 2023. 1

[37] Jiyuan Liu, Xinwang Liu, Siqi Wang, Xinhang Wan, Dongsheng Li, Kai Lu, and Kunlun He. Communication-efficient federated multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):17–32, 2026.

[38] Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *Science China Information Sciences*, 66(5):151101, 2023. 1

[39] Xuan Liu and Xiaobin Chang. Lora subtraction for drift-resistant space in exemplar-free continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15308–15318, 2025. 3, 6, 7

[40] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2023. 1, 3

[41] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995. 1

[42] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 1

[43] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):76–87, 2024. 1

[44] 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. 3

[45] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 2, 5

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[47] Zhuang Qi, Lei Meng, Zhaochuan Li, Han Hu, and Xiangxu Meng. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19986–19994, 2025. 1

[48] 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. 1

[49] 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 (CVPR)*, pages 2001–2010, 2017. 6

[50] Wuxuan Shi and Mang Ye. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1772–1781, 2023. 3

[51] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11909–11919, 2023. 6, 7

[52] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3097–3107, 2024. 1

[53] Hai-Long Sun, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Pilot: A pre-trained model-based continual learning toolbox. *arXiv preprint arXiv:2309.07117*, 2023. 6

[54] Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. Mos: Model surgery for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 20699–20707, 2025. 1

[55] Tianjiao Wan, Zijian Gao, Xudong Gong, Dawei Feng, Xingxing Zhang, Bo Ding, Yijie Wang, Huaimin Wang, and Kele Xu. Dynamic confidence variance for generalized coreset in active learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(3):3231–3245, 2026. 1

[56] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 398–414, 2022. 6

[57] 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. 1

[58] Tianqi Wang, Jingcai Guo, Depeng Li, and Zhi Chen. On the discrimination and consistency for exemplar-free class incremental learning. *arXiv preprint arXiv:2501.15454*, 2025. 3

[59] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648. Springer, 2022. 6, 7

[60] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pages 631–648. Springer, 2022. 1, 2, 3

[61] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 139–149, 2022. 6, 7

[62] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 1, 2, 3

[63] Kun Wei, Zhe Xu, and Cheng Deng. Compress to one point: Neural collapse for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21465–21473, 2025. 2, 6, 7

[64] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 1

[65] Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. *arXiv preprint arXiv:2501.13198*, 2025. 3, 6, 7

[66] 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 (CVPR)*, pages 3013–3022, 2021. 6

[67] 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. 1

[68] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6982–6991, 2020. 2

[69] Dulan Zhou, Zijian Gao, and Kele Xu. Perturbing to preserve: Defending fragile knowledge in online continual learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):28937–28945, 2026. 1

[70] 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 (ICLR)*, 2023. 6

[71] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision (IJCV)*, 2024. 6, 7

[72] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 23554–23564, 2024. 6, 7

[73] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23554–23564, 2024. 2

[74] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):9851–9873, 2024. 6

[75] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3): 1012–1032, 2025. 2, 6, 7

[76] Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[77] Fei Zhu, Zhen Cheng, Xu-yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14306–14318, 2021. 3

[78] 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 (CVPR)*, pages 5871–5880, 2021.

[79] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9296–9305, 2022. 3

[80] Tianyu Zhu, Hesong Li, and Ying Fu. Trim-sod: A multi-modal, multi-task, and multi-scale spacecraft optical dataset. *Space: Science Technology*, 5:0299, 2025. 1

[81] Huiping Zhuang, Run He, Kai Tong, Ziqian Zeng, Cen Chen, and Zhiping Lin. DS-AL: A dual-stream analytic learning for exemplar-free class-incremental learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):17237–17244, 2024. 6, 7