Supplementary Material of 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

1. Detailed Proofs of Theorem 1

In **Section 4.1** of the main paper, we introduced Theorem 1 to establish a tractable upper bound for the separation metric D_t. We now present the detailed theoretical proof of this theorem. To facilitate the proof, we first introduce several auxiliary lemmas, detailed as follows.

Lemma 1 (Upper Bound on Expected OOD Score)**.** *Let $P^{OOD}_{t,c'}$ be the feature distribution for an OOD class c'. Let $d_{M,t}(\cdot,\cdot)$ be the Mahalanobis distance w.r.t. Σ_t, and let $\alpha_{c,c'} := \frac{1}{2} d_{M,t}(\boldsymbol{\mu}_{IND,c}, \boldsymbol{\mu}_{OOD,c'})$. The expected score for a random OOD sample is bounded by:*

$$\mathbb{E}_{x\sim\bigcup_{m\neq t}\mathcal{D}_m}[ES_t(x)] \leq \frac{1}{k_{OOD}} \sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m} \sum_{c\in\mathcal{C}_t} \left((1 - P^{OOD}_{t,c'}(B_{\alpha_{c,c'}}(\boldsymbol{\mu}_{OOD,c'}))) + \exp\left(-\frac{\alpha^2_{c,c'}}{2}\right)\right) \tag{1}$$

where $B_r(u)$ is an open ball of radius r centered at u in the Mahalanobis distance.

Proof. The total expectation over OOD data is the average of expectations over each OOD class c', assuming uniform weighting:

$$\mathbb{E}_{x\sim\bigcup_{m\neq t}\mathcal{D}_m}[ES_t(x)] = \frac{1}{k_{\text{OOD}}} \sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m} \mathbb{E}_{x\sim\mathcal{D}_{m,c'}}[ES_t(x)].$$

By linearity, the inner expectation is a sum over the IND classes:

$$\mathbb{E}_{x\sim\mathcal{D}_{m,c'}}[ES_t(x)] = \sum_{c\in\mathcal{C}_t} \mathbb{E}_{z\sim P^{\text{OOD}}_{t,c'}}\left[\exp\left(-\frac{1}{2} d_{M,t}(z,\boldsymbol{\mu}_{\text{IND},c})^2\right)\right].$$

We analyze a single term from the sum. The expectation can be written as an integral split into the ball $B_{\alpha_{c,c'}}(\boldsymbol{\mu}_{\text{OOD},c'})$ and its complement $B^c_{\alpha_{c,c'}}$:

$$\begin{aligned}
&\mathbb{E}_{z\sim P^{\text{OOD}}_{t,c'}}\left[\exp\left(-\frac{1}{2} d_{M,t}(z,\boldsymbol{\mu}_{\text{IND},c})^2\right)\right] \\
&= \int_{B_{\alpha_{c,c'}}} \exp\left(-\frac{1}{2} d_{M,t}(z,\boldsymbol{\mu}_{\text{IND},c})^2\right) p^{\text{OOD}}_{t,c'}(z)dz \\
&\quad + \int_{B^c_{\alpha_{c,c'}}} \exp\left(-\frac{1}{2} d_{M,t}(z,\boldsymbol{\mu}_{\text{IND},c})^2\right) p^{\text{OOD}}_{t,c'}(z)dz.
\end{aligned}$$

To bound the first integral, consider any feature $z \in B_{\alpha_{c,c'}}(\boldsymbol{\mu}_{\text{OOD},c'})$. By definition, this implies $d_{M,t}(z,\boldsymbol{\mu}_{\text{OOD},c'}) \leq \alpha_{c,c'}$. The triangle inequality states $d_{M,t}(a,b) \leq d_{M,t}(a,c) + d_{M,t}(c,b)$. Applying this gives:

$$d_{M,t}(\boldsymbol{\mu}_{\text{IND},c}, \boldsymbol{\mu}_{\text{OOD},c'}) \leq d_{M,t}(\boldsymbol{\mu}_{\text{IND},c}, z) + d_{M,t}(z, \boldsymbol{\mu}_{\text{OOD},c'}).$$

Rearranging yields a lower bound for the distance of z to the IND prototype:

$$\begin{aligned}
d_{M,t}(z,\boldsymbol{\mu}_{\text{IND},c}) &\geq d_{M,t}(\boldsymbol{\mu}_{\text{IND},c}, \boldsymbol{\mu}_{\text{OOD},c'}) - d_{M,t}(z,\boldsymbol{\mu}_{\text{OOD},c'}) \\
&\geq 2\alpha_{c,c'} - \alpha_{c,c'} = \alpha_{c,c'}.
\end{aligned}$$

Using this lower bound for the distance and the fact that $\exp(\cdot) \leq 1$, we can bound the expectation:

$$\begin{aligned}
&\mathbb{E}_{z\sim P^{\text{OOD}}_{t,c'}}\left[\exp\left(-\frac{1}{2} d_{M,t}(z,\boldsymbol{\mu}_{\text{IND},c})^2\right)\right] \\
&\leq \int_{B_{\alpha_{c,c'}}} \exp\left(-\frac{\alpha^2_{c,c'}}{2}\right) p^{\text{OOD}}_{t,c'}(z)dz + \int_{B^c_{\alpha_{c,c'}}} 1 \cdot p^{\text{OOD}}_{t,c'}(z)dz \\
&= \exp\left(-\frac{\alpha^2_{c,c'}}{2}\right) P^{\text{OOD}}_{t,c'}(B_{\alpha_{c,c'}}) \\
&\quad + \left(1 - P^{\text{OOD}}_{t,c'}(B_{\alpha_{c,c'}})\right).
\end{aligned}$$

Using the simpler upper bound where $\exp(-\alpha^2/2) \cdot$

*Corresponding author. † Equal Contribution.

$P(B) \leq \exp(-\alpha^2/2)$ since $P(B) \leq 1$:

$$\mathbb{E}_{z\sim P^{\text{OOD}}_{t,c'}}\left[\exp\left(-\frac{1}{2}d_{M,t}(z,\boldsymbol{\mu}_{\text{IND},c})^2\right)\right] \leq \exp\left(-\frac{\alpha^2_{c,c'}}{2}\right) + \left(1 - P^{\text{OOD}}_{t,c'}(B_{\alpha_{c,c'}})\right).$$

Now, we substitute this bound back into the sum over all IND classes $c \in \mathcal{C}_t$:

$$\mathbb{E}_{x\sim\mathcal{D}_{m,c'}}[ES_t(x)] \leq \sum_{c\in\mathcal{C}_t}\left(\exp\left(-\frac{\alpha^2_{c,c'}}{2}\right) + \left(1 - P^{\text{OOD}}_{t,c'}(B_{\alpha_{c,c'}})\right)\right).$$

Finally, substituting this result into the average over all OOD classes c' yields the bound stated in the lemma:

$$\mathbb{E}_{x\sim\bigcup_{m\neq t}\mathcal{D}_m}[ES_t(x)] \leq \frac{1}{k_{\text{OOD}}}\sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m}\sum_{c\in\mathcal{C}_t}\left((1 - P^{\text{OOD}}_{t,c'}(B_{\alpha_{c,c'}}(\boldsymbol{\mu}_{\text{OOD},c'}))) + \exp\left(-\frac{\alpha^2_{c,c'}}{2}\right)\right).$$

This completes the proof.

Lemma 2 (Total Variation)**.** *Let $P_1, P_2 \in \mathcal{P}(\mathcal{X})$. The Total Variation is defined as:*

$$\delta(P_1, P_2) = \sup_{A\in\mathcal{B}} |P_1(A) - P_2(A)|, \tag{2}$$

where $\mathcal{B}$ denotes the Borel σ-algebra on $\mathcal{X}$. Furthermore, let $\mathcal{F}$ denote the unit ball in $L^\infty(\mathcal{X})$,

$$\mathcal{F} := \{f \in L^\infty(\mathcal{X}) \mid \|f\|_\infty \leq 1\}.$$

Then, the total variation distance can be equivalently characterized as:

$$\delta(P_1, P_2) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{x\sim P_1} f(x) - \mathbb{E}_{x\sim P_2} f(x)|. \tag{3}$$

Immediately following this, we recall that the Kullback-Leibler (KL) divergence for the multivariate Gaussian distribution.

Lemma 3 (Kullback-Leibler Divergence)**.** *Let $P_1 \sim \mathcal{N}(\mu_1, \Sigma)$ and $P_2 \sim \mathcal{N}(\mu_2, \Sigma)$, then we have the following:*

$$KL(P_1\|P_2) = \frac{1}{2}((\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)) = \frac{1}{2}d^2_M(\mu_1, \mu_2). \tag{4}$$

Next, we recall the following inequality that bounds the total variation by KL-divergence.

Lemma 4 (Pinsker's Inequality)**.** *Let $P_1, P_2 \in \mathcal{P}(\mathcal{X})$, then we have the following:*

$$\delta(P_1, P_2) \leq \sqrt{\frac{1}{2}KL(P_1\|P_2)}. \tag{5}$$

Furthermore, the subsequent version, which holds true when KL is large, is also valid:

$$\delta(P_1, P_2) \leq 1 - \frac{1}{2}\exp\left(-KL(P_1\|P_2)\right). \tag{6}$$

Lemma 5 (Upper Bound on the Separation Metric D_t)**.** *We define the scaled prototype distance as $\alpha_{c,c'} := \frac{1}{2}d_{M,t}(\boldsymbol{\mu}_{IND,c}, \boldsymbol{\mu}_{OOD,c'})$. The separation metric D_t defined in Eq. (6) is then bounded as follows:*

$$D_t \leq \frac{1}{k_{OOD}}\sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m}\sum_{c\in\mathcal{C}_t}\alpha_{c,c'}. \tag{7}$$

Proof of Lemma 5. First, by its definition (Eq. (6)), the score function $ES_t(x)$ is a sum of $C_t = k_{\text{IND}}$ terms, each bounded in $[0, 1]$. This implies $ES_t(x) \in [0, C_t]$.

Therefore, by Lemma 2, we have,

$$D_t = \mathbb{E}_{x\sim P_{\text{IND}}}[ES_t(x)] - \mathbb{E}_{x\sim P_{\text{OOD}}}[ES_t(x)] \leq C_t \cdot \delta(P_{\text{IND}}, P_{\text{OOD}}).$$

Next, we decompose the total variation distance. Let $P_{\text{IND}} = \frac{1}{C_t}\sum_{c\in\mathcal{C}_t} P_{\text{IND},c}$ and $P_{\text{OOD}} = \frac{1}{k_{\text{OOD}}}\sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m} P_{\text{OOD},c'}$. By the definition of total variation and the triangle inequality, the distance can be decomposed as follows:

$$\begin{aligned}
&\delta(P_{\text{IND}}, P_{\text{OOD}}) \\
&= \sup_{A\subseteq\mathcal{B}}\left|\frac{1}{C_t}\sum_{c\in\mathcal{C}_t} P_{\text{IND},c}(A) - \frac{1}{k_{\text{OOD}}}\sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m} P_{\text{OOD},c'}(A)\right| \\
&= \sup_{A\subseteq\mathcal{B}}\left|\frac{1}{C_t k_{\text{OOD}}}\sum_{c\in\mathcal{C}_t}\sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m}(P_{\text{IND},c}(A) - P_{\text{OOD},c'}(A))\right| \\
&\leq \frac{1}{C_t k_{\text{OOD}}}\sum_{c\in\mathcal{C}_t}\sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m}\sup_{A\subseteq\mathcal{B}}|P_{\text{IND},c}(A) - P_{\text{OOD},c'}(A)| \\
&= \frac{1}{C_t k_{\text{OOD}}}\sum_{c\in\mathcal{C}_t}\sum_{c'\in\bigcup_{m\neq t}\mathcal{C}_m}\delta(P_{\text{IND},c}, P_{\text{OOD},c'}).
\end{aligned}$$

By Lemma 4 and 3, for Gaussian feature distributions $P_{\text{IND},c} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{IND},c}, \Sigma_t)$ and $P_{\text{OOD},c'} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{OOD},c'}, \Sigma_t)$ with a shared covariance, we can bound the pairwise total

variation by the Mahalanobis distance:

$$\begin{aligned}\delta(P_{\text{IND},c}, P_{\text{OOD},c'}) &\leq \sqrt{\frac{1}{2} KL(P_{\text{IND},c} \| P_{\text{OOD},c'})} \\ &= \sqrt{\frac{1}{2} \cdot \frac{1}{2} d^2_{M,t}(\boldsymbol{\mu}_{\text{IND},c}, \boldsymbol{\mu}_{\text{OOD},c'})} \\ &= \frac{1}{2} d_{M,t}(\boldsymbol{\mu}_{\text{IND},c}, \boldsymbol{\mu}_{\text{OOD},c'}).\end{aligned}$$

Putting all together, we have

$$\begin{aligned}D_t &\leq C_t \cdot \left(\frac{1}{C_t k_{\text{OOD}}} \sum_{c \in \mathcal{C}_t} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \delta(P_{\text{IND},c}, P_{\text{OOD},c'}) \right) \\ &\leq \frac{1}{k_{\text{OOD}}} \sum_{c \in \mathcal{C}_t} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \frac{1}{2} d_{M,t}(\boldsymbol{\mu}_{\text{IND},c}, \boldsymbol{\mu}_{\text{OOD},c'}) \\ &= \frac{1}{k_{\text{OOD}}} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} \alpha_{c,c'}.\end{aligned}$$

This completes the proof.

Equipped with these lemmas, we now present the formal proof of Theorem 1.

Proof of Theorem 1. By Lemma 5, we obtain an upper bound for D_t as:

$$D_t \leq \frac{1}{2k_{\text{OOD}}} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c}, \mu_{\text{OOD},c'}).$$

Let

$$c_0(c') = \arg\min_{c \in \mathcal{C}_t} d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c}) \tag{8}$$

where $\mu_{\text{IND},c_0(c')}$ is the IND mean colsest to $\mu_{\text{OOD},c'}$. By the triangle inequality, we have:

$$\begin{aligned}d_M(\mu_{\text{IND},c}, \mu_{\text{OOD},c'}) \leq\ & d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \\ &+ d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}).\end{aligned} \tag{9}$$

Summing over all terms, we obtain:

$$\begin{aligned}\sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c}, \mu_{\text{OOD},c'}) \leq\ & \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \\ &+ \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}).\end{aligned}$$

Factoring out the constant terms, we have:

$$\begin{aligned}D_t &\leq \frac{1}{2k_{\text{OOD}}} \left(\sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \right. \\ &\quad \left. + \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}) \right) \\ &= \frac{1}{2k_{\text{OOD}}} \left(\sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \left(|\mathcal{C}_t| \cdot d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \right) \right. \\ &\quad \left. + \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}) \right) \\ &= \frac{1}{2k_{\text{OOD}}} \left(k_{\text{IND}} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \right. \\ &\quad \left. + |\bigcup_{m \neq t} \mathcal{C}_m| \cdot \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}) \right) \\ &= \frac{1}{2k_{\text{OOD}}} \left(k_{\text{IND}} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \right. \\ &\quad \left. + k_{\text{OOD}} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}) \right) \\ &= \frac{k_{\text{IND}}}{2k_{\text{OOD}}} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \\ &\quad + \frac{k_{\text{OOD}}}{2k_{\text{OOD}}} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}) \\ &= \frac{k_{\text{IND}}}{2k_{\text{OOD}}} \sum_{c' \in \bigcup_{m \neq t} \mathcal{C}_m} d_M(\mu_{\text{OOD},c'}, \mu_{\text{IND},c_0(c')}) \\ &\quad + \frac{1}{2} \sum_{c \in \mathcal{C}_t} d_M(\mu_{\text{IND},c_0(c')}, \mu_{\text{IND},c}).\end{aligned}$$

This completes the proof.

The derivation above establishes that the upper bound of D_t relies on the interplay between prototype separation and feature cohesion. This theoretically substantiates that minimizing the bound requires maximizing inter-class separation (**Hypothesis 1**) while enforcing intra-class compactness (**Hypothesis 2**).

2. Details on ETF Construction and Neural Collapse Properties

This section supplements the discussion in **Section 5.1** of the main text. Here, we provide the technical specifications for our ETF-based classifier and formally demonstrate its alignment with the NC phenomenon [4], specifically satisfying conditions NC2 through NC4.

NC Principles. The NC phenomenon [4] characterizes the terminal phase of training in deep classifiers through four distinct properties:

- **(NC1) Variability Collapse:** Within-class features collapse to their respective class means.
- **(NC2) ETF Convergence:** The class means (prototypes)

converge to a geometric structure known as an Equiangular Tight Frame (ETF).

- **(NC3) Self-Duality:** The classifier weights converge to align with the class means.
- **(NC4) NPC Convergence:** The decision boundaries converge to those of a Nearest Prototype Classifier (NPC).

Our GOD framework explicitly enforces (NC2), (NC3), and (NC4) by strictly fixing the classifier geometry, leaving only (NC1) to be optimized via our representation learning objective.

ETF Construction. As introduced in the main text, the fixed anchors E_t form a Simplex ETF. For a task with C_t classes in a feature space of dimension d (where $d \geq C_t - 1$), the weight matrix $E_t \in \mathbb{R}^{d \times C_t}$ is constructed as:

$$E_t = \sqrt{\frac{C_t}{C_t - 1}} U \left(I_{C_t} - \frac{1}{C_t} \mathbf{1}_{C_t} \mathbf{1}_{C_t}^T \right), \tag{10}$$

where I_{C_t} is the identity matrix, $\mathbf{1}_{C_t}$ is the all-ones vector, and $U \in \mathbb{R}^{d \times C_t}$ is a semi-orthogonal matrix satisfying $U^T U = I_{C_t}$, which embeds the simplex into the ambient feature space.

Properties. This construction guarantees: 1) *Maximal Separation (NC2):* The cosine similarity between any distinct pair of anchors is minimized and constant, specifically $\langle e_{t,i}, e_{t,j} \rangle = -\frac{1}{C_t - 1}$ for $i \neq j$. 2) *Zero-Mean (Part of NC3):* The anchors are centered at the origin, i.e., $\sum_{c=1}^{C_t} e_{t,c} = \mathbf{0}$.

Proof of NPC Equivalence (NC4). We demonstrate that maximizing cosine similarity (used in our framework) is mathematically equivalent to minimizing the Euclidean distance (standard NPC) when feature vectors are ℓ_2-normalized.

Let $z = z_t(x)$ and $e_c = e_{t,c}$ be unit-norm vectors (i.e., $\|z\|_2 = 1, \|e_c\|_2 = 1$). The classification decision rule proceeds as follows:

$$\begin{aligned}
\hat{y} &= \arg\min_c \|z - e_c\|_2^2 \\
&= \arg\min_c \left(\|z\|_2^2 + \|e_c\|_2^2 - 2\langle z, e_c \rangle \right) \\
&= \arg\min_c \left(1 + 1 - 2\langle z, e_c \rangle \right) \\
&= \arg\min_c \left(2 - 2\langle z, e_c \rangle \right) \\
&= \arg\max_c \langle z, e_c \rangle.
\end{aligned} \tag{11}$$

In summary, by hard-coding the ETF classifier, our framework structurally guarantees the properties of maximal separation (NC2), self-duality (NC3), and nearest-prototype decision boundaries (NC4) without requiring optimization. Consequently, the training process is relieved of the burden of learning a classifier, allowing it to focus entirely on the representation learning objective: compressing intra-class features to satisfy **NC1 (Variability Collapse)**. This focus directly minimizes the intra-class covariance, thereby tightening the generalization bound derived in our theoretical analysis.

3. Theoretical Guarantee of Deterministic OOD Separation

This section provides the rigorous theoretical substantiation for the claims made in **Section 5.1** of the main text. Specifically, we analyze the asymptotic behavior of the GOD detector under ideal training convergence.

Assumptions. Let t denote the current task and m denote any distinct task ($m \neq t$). **Recall from the main text** that our framework enforces a strict spherical geometry: all feature vectors $z(\cdot)$ and prototypes e are projected onto the unit hypersphere $\mathcal{S}^{d-1}$ (i.e., $\|z\|_2 = 1$ and $\|e\|_2 = 1$).

Based on this, we define the "Ideal Conditions" for our analysis:

1. **Ideal Intra-Task Collapse (IND Samples):** For any sample x belonging to class c of task t, the feature representation collapses perfectly to its class prototype:

$$\forall c \in \mathcal{C}_t,\ \forall x \in \mathcal{D}_{t,c} : \quad z_t(x) = e_{t,c}. \tag{12}$$

2. **Non-Alignment for OOD Samples:** For any sample x from the current task t, its representation in the projection space of another task m does not perfectly align with any prototype of task m:

$$\forall x \in \mathcal{D}_t,\ \forall k \in \mathcal{C}_m : \quad z_m(x) \neq e_{m,k}. \tag{13}$$

Since both $z_m(x)$ and $e_{m,k}$ are unit vectors, "alignment" refers strictly to directional identity. In high-dimensional spaces ($d \gg 1$), the probability of a sample from one distribution coincidentally having a cosine similarity of exactly 1.0 with a disjoint distribution's prototype is measure zero.

Proposition 1 (Deterministic Separation). Under the formulated ideal conditions, for any sample $x \in \mathcal{D}_{t,c}$, the GOD detector strictly separates the correct IND class logit from all possible OOD logits. Formally:

$$s_{t,c}(x) = \max_{j \in \mathcal{C}_t} s_{t,j}(x) > \max_{k \in \mathcal{C}_m} s_{m,k}(x), \quad \forall m \neq t. \tag{14}$$

Proof. Consider an arbitrary sample $x \in \mathcal{D}_{t,c}$.

1. In-Distribution Analysis (Task t): By Assumption 1, we have $z_t(x) = e_{t,c}$. Since we are on the unit hypersphere ($\|e_{t,c}\|_2 = 1$), the logits correspond to cosine similarities:

$$s_{t,c}(x) = \langle e_{t,c}, e_{t,c} \rangle = \|e_{t,c}\|_2^2 = 1, \tag{15}$$

$$s_{t,j}(x) = \langle e_{t,c}, e_{t,j} \rangle = -\frac{1}{C_t - 1} \quad (\forall j \neq c). \tag{16}$$

Since $C_t \geq 2$, we have $1 > -\frac{1}{C_t - 1}$. Thus, the maximum IND score is strictly 1:

$$\max_{j \in \mathcal{C}_t} s_{t,j}(x) = s_{t,c}(x) = 1. \tag{17}$$

Algorithm 1 **G**eometry-driven **O**OD **D**etectors (GOD) Training Procedure

Input: Sequential datasets $\{\mathcal{D}_t\}_{t=1}^{T}$; Pre-trained ViT backbone Φ; Hyperparameters: RP dimension d_B, ArcFace margin m, scaling factor s, EMA momentum α, and shared layer count Num_{SL}.
Initialize: Shared Random Projection RP (Fixed, mapping to $\mathbb{R}^{d_B}$).
for task $t = 1, \ldots, T$ **do**
  ## 1. Hybrid Architecture & Geometry Initialization
  if $t = 1$ **then**
    Construct ETF anchors $E_t = \{e_{t,1}, \ldots, e_{t,C_t}\}$ satisfying the ETF property (Shared geometry).
    Initialize Shared LoRA module θ_{SL} (defined by Num_{SL}).
    Initialize Task-specific LoRA θ_{TL_1} and Projection Head θ_{TP_1}.
    Set learnable parameters $\Theta_t = \{\theta_{\text{SL}}, \theta_{\text{TL}_1}, \theta_{\text{TP}_1}\}$.
  else
    Retrieve frozen θ_{SL}, RP, and shared anchors E_t.
    Initialize θ_{TL_t} with weights from $\theta_{\text{TL}_{t-1}}$ (Warm Start).
    Initialize θ_{TP_t} randomly.
    Set learnable parameters $\Theta_t = \{\theta_{\text{TL}_t}, \theta_{\text{TP}_t}\}$.
  end if
  ## 2. Geometry-driven Optimization
  while not converged **do**
    Sample batch (x, y) from $\mathcal{D}_t$.
    Extract features via Hybrid LoRA: $f_t(x) = \Phi(x; \theta_{\text{SL}}, \theta_{\text{TL}_t})$.
    Project to unit hypersphere:
      $z_t(x) = \text{Normalize}(\mathrm{TP}_t(\mathrm{RP}(f_t(x))))$.
    Compute logits via shared ETF anchors: $s_{t,c}(x) = \langle z_t(x), e_{t,c} \rangle$.
    Compute Separation Loss $\mathcal{L}_{\text{etf}}$ via Eq. (14).
    Compute Compactness Loss $\mathcal{L}_{\text{arc}}$ via Eq. (15).
    Update Θ_t to minimize $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{etf}} + \mathcal{L}_{\text{arc}}$.
  end while
  Save Θ_t for inference (Anchors E_t are shared and persistent).
end for

Algorithm 2 Dual-Mode Inference Strategy

Input: Test sample x; Universal EMA Adapter Θ^{ema}; Task-specific parameters $\{\Theta_t\}_{t=1}^{T}$; Shared parameters $\theta_{\text{SL}}, \mathrm{RP}, E_t$; Top-$k$ parameter k.
Output: Predicted Label y_{pred}.

1. Coarse Mode (Fast, Single Forward Pass)
Extract shallow features via Shared LoRA: $h_{\text{SL}} = \Phi_{\text{shallow}}(x; \theta_{\text{SL}})$.
Extract deep features via EMA Adapter: $f^{\text{ema}}(x) = \Phi_{\text{deep}}(h_{\text{SL}}; \Theta^{\text{ema}})$.
Compute global logits: $\text{logits}^{\text{EMA}}(x) = \text{Concat}_{t=1}^{T}\left(\langle \text{Normalize}(\mathrm{TP}_t(\mathrm{RP}(f^{\text{ema}}(x)))), E_t \rangle\right)$.
Obtain Coarse Prediction: $\text{Label}_{\text{Coa}} = \arg\max \text{logits}^{\text{EMA}}(x)$.
2. Refined Mode (Optional, Sparse Activation)
if Mode == Refined **then**
  // Select candidate tasks based on Top-k classes
  Identify candidate tasks: $\mathcal{T}_{\text{top-}k} = \{\text{TaskID}(c) \mid c \in \text{Top-}k(\text{logits}^{\text{EMA}}(x))\}$.
  Initialize refined logits: $\text{logits}^{\text{Ref}}(x) = \emptyset$.
  for each task $\tau \in \mathcal{T}_{\text{top-}k}$ **do**
    // Re-activate specific experts using cached h_{SL}
    Extract specific features: $f_\tau(x) = \Phi_{\text{deep}}(h_{\text{SL}}; \Theta_\tau)$.
    Compute refined logits: $\text{logits}_\tau^{\text{Ref}}(x) = \langle \text{Normalize}(\mathrm{TP}_\tau(\mathrm{RP}(f_\tau(x)))), E_\tau \rangle$.
    Update: $\text{logits}^{\text{Ref}}(x) \leftarrow \text{logits}^{\text{Ref}}(x) \cup \text{logits}_\tau^{\text{Ref}}(x)$.
  end for
  Obtain Refined Prediction: $\text{Label}_{\text{Ref}} = \arg\max \text{logits}^{\text{Ref}}(x)$.
  return $\text{Label}_{\text{Ref}}$
else
  return $\text{Label}_{\text{Coa}}$
end if

2. Out-of-Distribution Analysis (Task $m \neq t$): Consider any class k from an OOD task m. The logit is $s_{m,k}(x) = \langle z_m(x), e_{m,k} \rangle$. Applying the Cauchy-Schwarz inequality:

$$\langle z_m(x), e_{m,k} \rangle \leq \|z_m(x)\|_2 \cdot \|e_{m,k}\|_2. \tag{18}$$

Using the unit norm constraint ($\|z\| = \|e\| = 1$), this bound simplifies to:

$$s_{m,k}(x) \leq 1. \tag{19}$$

Crucially, the equality in Cauchy-Schwarz holds if and only if the two vectors are collinear and in the same direction. By Assumption 2 ($z_m(x) \neq e_{m,k}$), perfect alignment is excluded. Therefore, the inequality is strict:

$$s_{m,k}(x) < 1 \implies \max_{k \in \mathcal{C}_m} s_{m,k}(x) < 1. \tag{20}$$

Conclusion: Combining Eq. (17) and Eq. (20), we obtain the strict inequality:

$$s_{t,c}(x) = 1 > \max_{k \in \mathcal{C}_m} s_{m,k}(x).$$

This concludes the proof.

Thus, this proof serves as the theoretical foundation for **Section 5.1**, confirming that under ideal conditions, the GOD framework inherently possesses a deterministic guarantee for perfect IND classification and OOD rejection.

Table 1. Comparison of Average Accuracy $\bar{A}(\%)$ and Last-Task Accuracy $A_T(\%)$ on ImageNet-R and Stanford Cars with $T = 20$ tasks. The best results among baselines are marked in blue, and the improvements of our method over them are marked in red.

Method	ImageNet-R $T = 20$		Stanford Cars $T = 20$	
	$\bar{A}$	A_T	$\bar{A}$	A_T
Finetune	71.14	63.72	46.92	33.39
LwF (TPAMI 2018) [2]	70.74	64.45	37.75	19.19
DualPrompt (ECCV 2022) [8]	73.61	67.12	45.09	29.68
L2P (CVPR 2022) [9]	75.59	68.73	49.86	39.59
CODA-Prompt (CVPR 2023) [6]	71.63	67.93	26.56	14.98
LAE (ICCV 2023) [1]	72.85	65.57	41.63	27.27
DS-AL (AAAI 2024) [17]	75.90	74.05	63.17	13.97
Aper (IJCV 2024) [14]	76.28	69.25	50.94	38.25
EASE (CVPR 2024) [15]	78.15	71.10	50.73	35.81
SimpleCIL (IJCV 2024) [16]	67.60	61.35	50.89	38.26
NC-CIPM (AAAI 2025) [10]	74.70	68.62	61.40	41.05
LORA-DRS (CVPR 2025) [3]	79.88	72.12	56.65	43.69
SD-LoRA (ICLR 2025) [11]	81.68	73.15	54.77	43.77
GOD (All)	**82.86**	**75.05**	**65.93**	**49.50**
GOD (Coarse)	**81.19**	**73.57**	**63.22**	**47.20**
GOD (Refined)	**82.88** (+1.20)	**75.60** (+1.55)	**65.15** (+1.98)	**48.93** (+5.16)

4. Pseudo Code

To facilitate reproducibility, we provide the detailed pseudocode for our Geometry-driven OOD Detectors (GOD) framework. **Algorithm 1** outlines the training procedure, explicitly detailing the construction of the shared ETF geometry and the optimization of the Parameter-Efficient Hybrid Architecture. It highlights the distinct handling of the initial task (learning shared components) versus incremental tasks (freezing shared modules and warm-starting task-specific ones) under the composite objective of separation ($\mathcal{L}_{\text{etf}}$) and compactness ($\mathcal{L}_{\text{arc}}$). **Algorithm 2** presents the Dual-Mode Inference strategy, demonstrating how the *Coarse Mode* utilizes the EMA adapter for rapid global prediction, and how the *Refined Mode* selectively activates specific experts based on Top-k candidates to achieve an optimal trade-off between accuracy and efficiency.

5. Implementation Details

Expanding on the experimental setup in **Section 6.1**, we provide further details regarding the network architecture and training strategy. The GOD framework employs a specific two-stage projection head comprising a shared Random Projection (RP) layer and a Task-specific Projection (TP_t) layer. The fixed RP layer projects 768-dimensional features into a 3000-dimensional space, while the TP_t layer maps them back to the 768-dimensional ETF space. To ensure smooth adaptation, TP_t is randomly initialized for the first task but inherits weights from the preceding task for all subsequent stages. Training is conducted for 40 epochs per task with a batch size of 48, utilizing an SGD optimizer (momentum 0.9) and a cosine annealing learning rate schedule starting at 0.01. For GOD-specific hyperparameters, the EMA momentum α is set to 0.99. In the ArcFace loss component, the angular margin m and scale factor s are set to 0.4 and 64, respectively.

Table 2. Comparison to traditional replay-based CIL methods.

Method	Instances	CIFAR-100 ($T = 10$)		Imagenet-R ($T = 10$)	
		$\bar{A}$	A_T	$\bar{A}$	A_T
iCaRL [5]	20 / classes	82.46	73.87	72.96	60.67
DER [12]	20 / classes	86.04	77.93	80.48	74.32
FOSTER [7]	20 / classes	89.87	84.91	81.34	74.48
MEMO [13]	20 / classes	84.08	75.79	74.80	66.62
GOD (Refined)	0	**92.87**	**88.11**	**85.41**	**79.20**

Table 3. Ablation study comparing the impact of using different components in GOD (Refined) on the average incremental accuracy $\bar{A}(\%)$.

$\mathcal{L}_{etf}$	$\mathcal{L}_{arc}$	RP	Stanford Cars		Imagenet-R	
			$T = 10$	5	$T = 10$	5
✓			51.86	57.99	80.10	82.99
✓	✓		65.55	75.74	84.40	86.31
✓		✓	65.99	74.37	85.07	86.83
✓	✓	✓	**77.10**	**85.23**	**85.41**	**86.90**

6. Extended Evaluation

Here we evaluate the robustness of GOD on a more challenging scenario with total 20 incremental tasks ($T = 20$). This setting significantly amplifies the risk of performance degration due to the increased task numbers. Table 1 presents the comparative results. GOD consistently outperforms state-of-the-art methods across these long sequences. Notably, on ImageNet-R, **GOD (Refined)** achieves a compelling average accuracy ($\bar{A}$) of 82.88%, surpassing the strongest baseline by 1.20%. Similarly, on Stanford Cars, our method demonstrates exceptional stability, maintaining a Last-Task Accuracy (A_T) of 48.93%—a substantial margin of +5.16% over the runner-up. These results empirically validate that our GOD effectively mitigates feature collapse even over prolonged training phases.

7. Comparison to Replay-based Methods

In Table 2, we compare GOD with competitive replay-based methods (storing 20 instances per class) using the same PTM. The results are excerpted from [15]. Even without saving instances, GOD shows substantial improvements on CIFAR-100 and ImageNet-R, outperforming the best baseline methods by 3.00% and 3.20% on $\bar{A}$ and A_T for CIFAR-100, and by 4.07% and 4.72% on $\bar{A}$ and A_T for ImageNet-R, achieving superior results with a clear margin and further proving its effectiveness.

8. Additional Quantitative Ablation Studies

Table 3 presents the quantitative ablation results corresponding to the visualizations discussed in the main text. We report the Average Incremental Accuracy ($\bar{A}$) across different benchmarks to validate that the observed geometric improvements translate into tangible performance gains. The baseline model, trained solely with $\mathcal{L}_{\text{etf}}$, suffers from

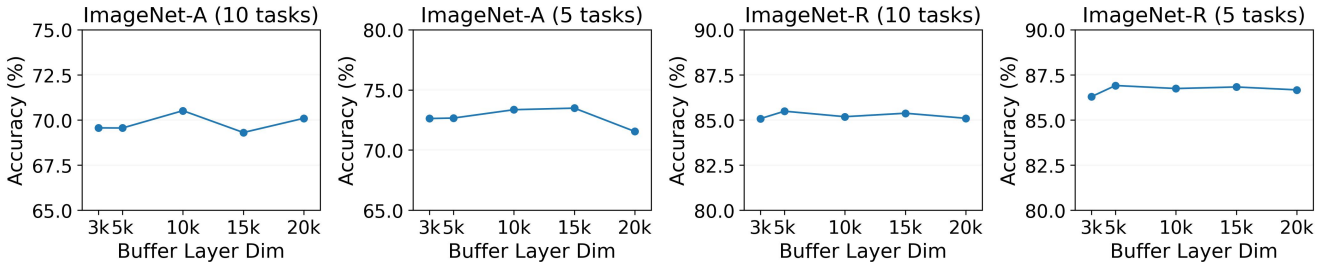

Figure 1. The impact of Random Projection dimension d_B on metric $\bar{A}(\%)$ in GOD (Refined).

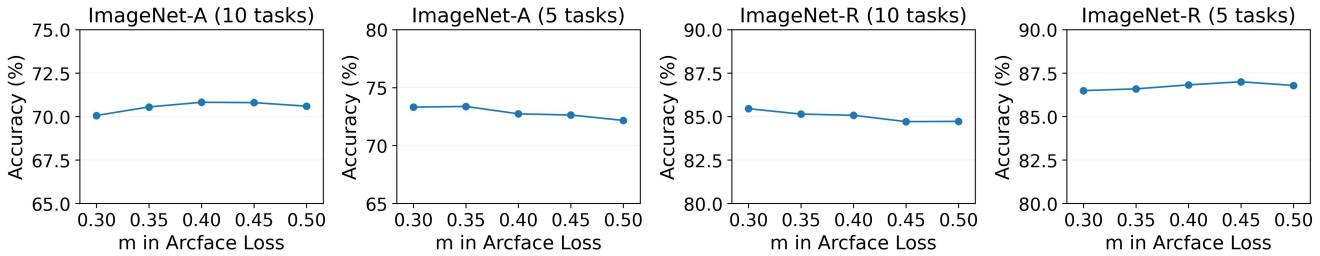

Figure 2. The impact of ArcFace margin m on metric $\bar{A}(\%)$ in GOD (Refined).

loose feature distributions, resulting in suboptimal performance. For instance, on the Stanford Cars dataset ($T = 10$), the baseline yields a modest $\bar{A}$ of 51.86%. Introducing individual components brings clear improvements: adding $\mathcal{L}_{\text{arc}}$ to enforce intra-class compactness boosts the accuracy significantly, while applying RP to enhance inter-class separation also yields consistent gains. Specifically, on ImageNet-R ($T = 10$), using $\mathcal{L}_{\text{arc}}$ alone improves the $\bar{A}$ from 80.10% to 84.40%, a gain of over 4.0%. Crucially, the full framework achieves the best performance across all settings, demonstrating the complementary nature of our proposed components. This synergy is particularly pronounced on the fine-grained Stanford Cars dataset ($T = 10$), where the combination achieves a remarkable $\bar{A}$ of 77.10%. This represents a substantial improvement of 11.55% compared to using $\mathcal{L}_{\text{arc}}$ alone (65.55%), validating that simultaneously enforcing "high intra-class cohesion" and "low inter-class coupling" is essential for robust incremental learning.

9. Hyperparameter Sensitivity Analysis

We conducted a comprehensive sensitivity analysis on four critical hyperparameters. The results consistently demonstrate that GOD maintains high robustness across varying configurations.

Random Projection Dimension (d_B). Figure 1 investigates the sensitivity of GOD (Refined) to the projection dimension d_B on the ImageNet-A and ImageNet-R benchmarks. The results demonstrate remarkable stability across a wide range of dimensions ($3k \sim 20k$). Specifically, performance on ImageNet-R remains essentially invariant to variations in d_B, while ImageNet-A exhibits only marginal fluctuations. Based on these observations, we adopt $d_B = 3000$ as the default configuration. This choice minimizes the computational footprint without compromising accuracy, as evidenced by the fact that the model maintains peak performance levels (e.g., $> 85\%$ on ImageNet-R) even at this compact dimensionality, significantly outperforming the comparison methods.

Table 4. Average incremental accuracy $\bar{A}(\%)$ comparison of different values of the number of shared layers Num_{SL} across various benchmarks and settings in GOD (Refined).

Num_{SL}	Stanford Cars		Imagenet-R	
	$T = 10$	5	$T = 10$	5
0	**77.36**	85.00	85.33	**86.99**
3	77.33	**85.29**	85.34	86.75
6	77.24	85.05	85.16	86.79
9	77.10	85.23	**85.41**	86.90

ArcFace Margin (m). Consistent with our findings on projection dimension, the margin parameter m demonstrates exceptional stability. As visualized in Figure 2, the accuracy curves across both ImageNet-A and ImageNet-R benchmarks are remarkably flat over the investigated range of $[0.30, 0.50]$. The performance fluctuations are negligible; for instance, on ImageNet-R ($T = 10$), the variation in accuracy is less than 0.5%. Given this insensitivity, we select $m = 0.4$ as the default configuration. Crucially, with this fixed margin, our method consistently delivers superior performance compared to the state-of-the-art methods on both ImageNet-A and ImageNet-R benchmarks.

Number of Shared Layers (Num_{SL}). Turning to the architectural design, Table 4 examines the performance stability under different parameter sharing configurations.

Table 5. Average incremental accuracy $\bar{A}(\%)$ comparison of different values of EMA momentum α across various benchmarks and settings in GOD (Refined).

α	Stanford Cars		Imagenet-R	
	$T = 10$	5	$T = 10$	5
0.9	**77.38**	85.23	84.27	86.46
0.95	77.28	85.08	84.62	86.64
0.99	77.10	85.23	**85.41**	**86.90**
0.995	77.31	85.31	85.29	86.78
0.999	77.04	**85.41**	84.94	86.67

The proposed method exhibits remarkable robustness to the extent of feature sharing, with fluctuations in Average Incremental Accuracy ($\bar{A}$) remaining negligible (e.g., $\sim 0.3\%$) between independent learning ($Num_{SL} = 0$) and deep sharing ($Num_{SL} = 9$). Consequently, we adopt $Num_{SL} = 9$ as the default setting to maximize parameter efficiency. Importantly, even under this highly efficient configuration, our method maintains superior performance over the state-of-the-art approaches. For instance, on the Stanford Cars dataset ($T = 10$), the setting of $Num_{SL} = 9$ yields an $\bar{A}$ of 77.10%, which still significantly outperforms the previous best result (SD-LoRA: 73.05%) by over 4%. This confirms that our framework achieves an optimal balance between architectural efficiency and classification accuracy.

EMA Momentum (α). Finally, Table 5 analyzes the impact of the EMA momentum coefficient α. The results demonstrate that our framework is highly robust to variations in this hyperparameter, with the **Average Incremental Accuracy ($\bar{A}$)** exhibiting minimal fluctuations (mostly $< 1\%$) across the broad range of $[0.9, 0.999]$. Based on these findings, we adopt $\alpha = 0.99$ as the default configuration. Theoretically, a higher momentum (closer to 1) slows down the evolution of the teacher model, thereby prioritizing feature stability and effectively mitigating catastrophic forgetting. Crucially, regardless of the specific α value chosen within this range, our method consistently yields results superior to the state-of-the-art. For instance, on ImageNet-R ($T = 10$), even with the fixed setting of $\alpha = 0.99$, we achieve an $\bar{A}$ of 85.23%, which remains significantly higher than the previous best result of 82.88% (SD-LoRA), further validating the reliability and effectiveness of our approach.

References

[1] 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

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

[3] 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. 6

[4] 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. 3

[5] 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

[6] 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] 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

[8] 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

[9] 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

[10] 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. 6

[11] 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. 6

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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. 6

[17] 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