Geometry-Anchored Transport Framework for Exemplar-Free Class-Incremental Learning

Hongye Xu[1] and Bartosz Krawczyk[1]

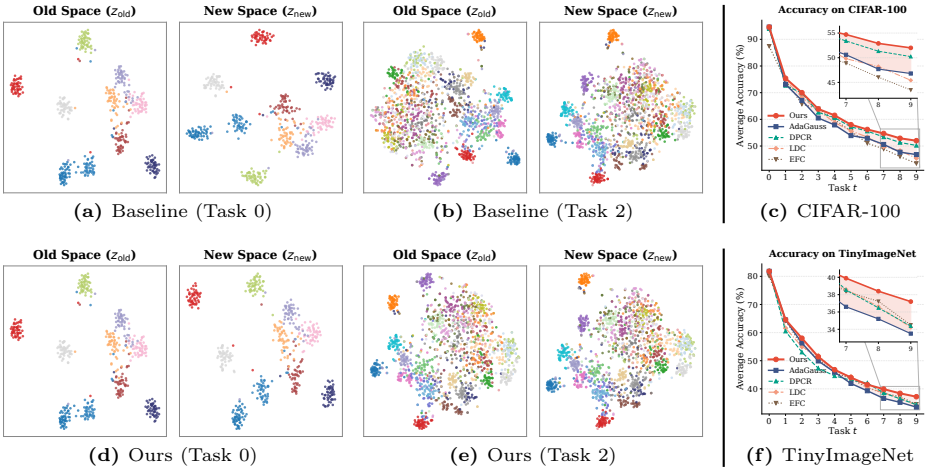
Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623, USA
{hx5239,bartosz.krawczyk}@rit.edu

Abstract. Exemplar-free class-incremental learning (EFCIL) requires stable decision boundaries within a shifting feature space. While maintaining class-conditional Gaussian statistics provides a principled classification strategy, these parametric summaries remain sensitive to *anisotropic representation drift*. Existing methods often transport these statistics across tasks using a decoupled, post-hoc paradigm: optimizing a backbone without explicit geometric constraints can distort the legacy manifold, limiting the precision of retroactive alignment. In this paper, we formulate feature transport as an endogenous training constraint rather than a separate post-task step, presenting the **Geometry-Anchored Transport Framework**. First, we derive an *Analytic Geometric Anchor* via Mahalanobis-aligned regression to mitigate macroscopic anisotropic drift. Second, we introduce a *Topology-Aware Evolution* objective that regularizes localized manifold degradation while calibrating a residual network against the analytic prior. By coupling manifold evolution with transport constraints during the primary training phase, our framework mitigates evaluation errors without requiring decoupled fine-tuning. Experiments across CIFAR-100, TinyImageNet, and ImageNet-100 demonstrate that the proposed framework consistently improves upon existing post-hoc alternatives under strict exemplar-free constraints. The code is available at `https://github.com/HXuSz11/GATF_ECCV2026`.

Keywords: Continual Learning · Class-Incremental Learning · Exemplar-Free

1 Introduction

The primary challenge in Exemplar-Free Class-Incremental Learning (EFCIL) is managing the plasticity-stability trade-off within a non-stationary feature space [9, 13, 19, 27, 34]. Without the safety net of memory buffers [1], models are vulnerable to catastrophic forgetting [2, 8, 26, 30]. To address this, recent methods have increasingly adopted prototype-based and Gaussian-parametric inference [6, 7, 23, 28, 29]. By summarizing previously seen distributions into compact sufficient statistics [22], networks can utilize Bayes or Mahalanobis decision rules across the sequence of learned tasks [6, 29].



(a) Baseline (Task 0) **(b)** Baseline (Task 2) **(c)** CIFAR-100

(d) Ours (Task 0) **(e)** Ours (Task 2) **(f)** TinyImageNet

Fig. 1: Topological anchoring and continual learning performance. (Left & Middle): Feature transport from the frozen model (z_{old}) to the new model (z_{new}) visualized on TinyImageNet. In the baseline (AdaGauss [29]) (**Top**), the embedding space undergoes significant rotation and non-uniform displacement, leading to a disruption of the relative spatial relationships between categories. Conversely, our framework (**Bottom**) anchors the global topology, effectively mitigating these geometric distortions to maintain structural alignment across tasks. **(Right):** This spatial preservation corresponds to sustained accuracy, improving upon established baselines. A detailed discussion regarding our baseline selection is provided in the SUPPLEMENTARY.

However, this parametric approach encounters the issue of *representation drift* [6, 12, 14, 23, 28, 29, 33, 38, 39]. As the backbone assimilates novel categories, the underlying embedding space undergoes continuous non-linear deformation. To update prior statistics, the standard paradigm employs a decoupled "train-then-adapt" sequence [5, 7, 11, 23]. In this formulation, the backbone is first updated on new data using knowledge distillation, and subsequently, a post-hoc adapter is trained to map legacy prototypes into the modified space.

This decoupled approach introduces two structural limitations. First, optimizing the backbone—driven primarily by the cross-entropy of new classes and standard distillation—can induce *topological degradation* [4, 9, 13]. Evolving the network without cross-space geometric constraints can distort the neighborhood structure of prior categories, complicating subsequent post-hoc mappings. Second, Mahalanobis-based evaluation is sensitive to *anisotropic drift*. Transport errors along low-variance directions are amplified by the inverse covariance matrix, which can erode decision margins and introduce classification bias between old and new classes.

To address these limitations, we propose the **Geometry-Anchored Transport Framework**, integrating feature transport as a geometry-aware constraint during representation learning rather than treating it solely as a post-hoc adjustment.

The framework relies on an **Analytic Geometric Anchor (AGA)**. Using a closed-form generalized least squares (GLS) prior aligned with the Mahalanobis geometry, we establish a baseline mapping to mitigate macroscopic anisotropic drift and suppress tail-end errors. However, applying a static analytic prior to an evolving neural network necessitates topological stability. Consequently, we structure the backbone's training through *Topology-Aware Evolution*. By incorporating a local anchoring constraint directly into the optimization trajectory, we restrict the backbone from parameter updates that degrade legacy neighborhood structures. This encourages smoother backbone evolution, allowing a residual network to calibrate against the non-linear deformations not captured by the analytic prior, thereby reducing the reliance on decoupled post-hoc fine-tuning.

Contributions. Our contributions are summarized as follows:

- We analyze the limitations of the decoupled post-hoc transport paradigm, showing that topological degradation and Mahalanobis-amplified anisotropic drift contribute to margin erosion in EFCIL.
- We introduce an Analytic Geometric Anchor formulated via a Sylvester equation, designed to mitigate macroscopic geometric drift and bound low-variance transport errors.
- We propose a topology-aware evolution strategy that calibrates a non-linear residual against the analytic prior while grounding the active feature space. This coupled framework provides margin stability bounds (Theorem 2) and demonstrates competitive performance across CIFAR-100, TinyImageNet, ImageNet-100, and CUB-200.

2 Related Work

Geometry-Aware Continual Evaluation. To mitigate catastrophic forgetting without storing raw samples, recent research has focused on adapting the evaluation geometry. Since standard Euclidean metrics may degrade under representation drift, approaches such as FeCAM [6] utilize anisotropic, class-conditional covariances to construct geometry-aware decision boundaries. Similarly, FeTrIL [28] freezes the feature extractor early and synthesizes pseudo-features to maintain old-class separability. While these methods improve stability, their reliance on largely static feature spaces can restrict plasticity for new distributions. Our framework utilizes Mahalanobis evaluation but avoids the static-backbone constraint. Instead, we regularize the evaluation geometry during continuous training via an explicit analytic transport anchor.

Predictive Feature Mapping and Transport. To maintain plasticity while updating prior statistics, recent EFCIL methods map legacy representations into the active feature space. Initial approaches, such as SDC [33] and ADC [7], estimated global or class-wise translations. Subsequent methods, including LDC [5],

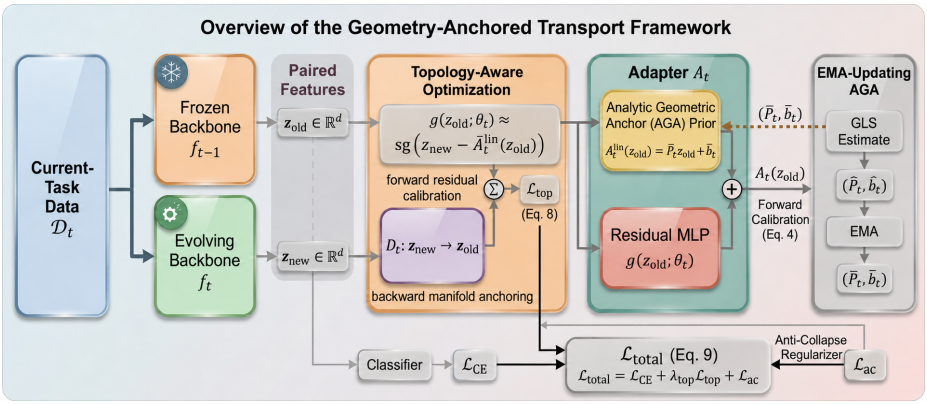


Fig. 2: Overview of the Geometry-Anchored Transport Framework. At task t, the active backbone f_t learns from the new dataset $\mathcal{D}_t$, while the frozen f_{t-1} provides legacy features z_{old}. Rather than relying solely on decoupled post-hoc adaptation, we integrate feature transport into the training phase. **Macroscopically**, an Analytic Geometric Anchor (AGA) computes a closed-form linear prior to mitigate anisotropic drift. **Microscopically**, the backbone utilizes Topology-Aware Evolution: a backward distiller D_t anchors the legacy manifold to restrict topological degradation, and a forward residual network g calibrates against the AGA's displacement. Ultimately, this coupled optimization yields a geometry-aware adapter A_t that facilitates the stable transport of legacy Gaussian statistics for robust Mahalanobis evaluation.

EFC [23], AdaGauss [29], and DPCR [11], employ auxiliary neural networks to project cached prototypes forward based on pairwise feature relationships.

Despite their strong empirical performance, these architectures often rely on a decoupled, post-hoc paradigm. The active backbone is optimized largely independently of the downstream transport mechanism, typically constrained by standard distillation. This separation implies that the post-hoc adapter must map features across a manifold that may undergo topological degradation during the primary optimization phase. Furthermore, auxiliary neural adapters typically optimize isotropic Euclidean objectives, which may under-penalize the low-variance anisotropic distortions that affect Bayes classifiers. Our methodology addresses this decoupled structure. By solving macroscopic drift analytically and embedding a topology-aware constraint directly into the backbone optimization, we constrain manifold evolution to bound transport errors during the primary training phase.

Relation to DPCR. DPCR [11] is the closest recent post-hoc alternative to our framework. It estimates semantic shift after backbone training through dual projection and reconstructs a classifier with ridge regression. In contrast, our method does not aim to repair the classifier after the representation has already changed. Instead, the AGA provides a Mahalanobis-aligned analytic prior during backbone optimization, and the topology-aware objective constrains feature evolution before old-class geometry is distorted. Thus, the central difference is not

merely the form of the transport map, but the timing of the constraint: DPCR performs post-training calibration, whereas our framework enforces training-time geometry anchoring to make the final statistic pushforward more reliable.

3 Methodology

3.1 Setup and Evaluation Space

We study an exemplar-free class-incremental learning problem with tasks indexed by $t \in \{0, 1, \dots\}$. After completing task $t-1$, we freeze the previous backbone f_{t-1} and train a new backbone f_t using only the current-task dataset $\mathcal{D}_t$. For an input x, we denote:

$$z_{\text{old}} = f_{t-1}(x) \in \mathbb{R}^d, \qquad z_{\text{new}} = f_t(x) \in \mathbb{R}^d. \tag{1}$$

Evaluation is performed in the active feature space induced by f_t using class-conditional Gaussian statistics. For newly introduced classes $c \in \mathcal{C}_t$, the moments (μ_c^t, Σ_c^t) are estimated directly from $\mathcal{D}_t$ under f_t. For old classes $c \in \mathcal{C}_{1:t-1}$, the archived statistics $(\mu_c^{t-1}, \Sigma_c^{t-1})$ remain untouched during task-t training. After training converges, we perform a training-free distribution pushforward:

$$\tilde{z}_m \sim \mathcal{N}(\mu_c^{t-1}, \Sigma_c^{t-1}), \qquad v_m = A_t(\tilde{z}_m), \tag{2}$$

and re-estimate $(\hat{\mu}_c^t, \hat{\Sigma}_c^t)$ from $\{v_m\}_{m=1}^M$ in the active feature space. We denote the assembled task-t statistics for all seen classes by (μ_c^t, Σ_c^t), where old classes use the transported estimates and new classes use empirical estimates.

Unless stated otherwise, inference uses a Gaussian Bayes/Mahalanobis rule over all seen classes:

$$s_c(x) = (z - \mu_c^t)^\top (\Sigma_c^t)^{-1} (z - \mu_c^t), \qquad \hat{y} = \arg\min_{c \in \mathcal{C}_{1:t}} s_c(x), \tag{3}$$

where $z = f_t(x)$.

To relate z_{old} and z_{new}, we incorporate a co-evolving adapter A_t. Unlike prior methods that relegate transport to a post-hoc stage, we treat A_t as an integral constraint during primary training, so that the feature space evolves in a form that remains compatible with the task-end pushforward of old-class Gaussian statistics.

3.2 Geometry-Anchored Transport Framework

Bridging two independently evolving high-dimensional spaces presents distinct optimization challenges. A conventional approach decouples feature learning from cross-space alignment, optimizing the backbone independently of legacy structures before fitting a post-hoc adapter. However, this decoupled paradigm can limit overall stability: if stable classification relies on bounded transport errors (Theorem 2), the backbone's evolutionary trajectory should be explicitly

constrained to preserve these bounds. Unconstrained optimization leaves the mapping vulnerable to anisotropic representation drift—where distortions along low-variance yet decision-critical directions can shift Mahalanobis margins.

To address this, our framework approaches the transport problem by introducing an explicit Analytic Geometric Anchor (AGA) to resolve macroscopic anisotropic drift analytically. This is supported by a Topology-Aware Evolution mechanism that regularizes the backbone's local deformations during the primary training phase.

The Analytic Geometric Anchor (AGA). To make old→new transport geometrically aligned, we anchor the forward adapter A_t with a closed-form linear prior, which we term the Analytic Geometric Anchor (AGA). We parameterize the forward transport as an explicitly decomposed function:

$$A_t(z) = A_t^{\text{lin}}(z) + g(z;\theta_t), \qquad A_t^{\text{lin}}(z) = P_t z + b_t, \tag{4}$$

where $P_t \in \mathbb{R}^{d\times d}$ and $b_t \in \mathbb{R}^d$ define an affine map representing the AGA, and $g(\cdot;\theta_t)$ is a small MLP learning localized residual corrections.

The affine component captures the dominant global drift, while the residual network models localized non-linear corrections. During training, the linear component is used as a zero-gradient prior: gradients are not propagated through the closed-form estimate, but the active backbone and residual network are optimized against it.

We estimate this linear relationship using paired current-task features $x \sim \mathcal{D}_t$, with $\mu_{\text{old}} = \mathbb{E}[z_{\text{old}}]$, $\mu_{\text{new}} = \mathbb{E}[z_{\text{new}}]$, and $C_{\text{on}} = \text{Cov}(z_{\text{old}}, z_{\text{new}})$. To target Mahalanobis-sensitive tail errors (empirically analyzed in Sec. 4.2), we formulate the estimation of P_t, b_t as a ridge-regularized Generalized Least Squares (GLS) problem:

$$\min_{P_t,b_t} \mathbb{E}\Big[\|(P_t z_{\text{old}} + b_t) - z_{\text{new}}\|^2_{\Sigma_{\text{new}}^{-1}}\Big] + \rho\|P_t\|_F^2. \tag{5}$$

Centering yields the optimal intercept $b_t = \mu_{\text{new}} - P_t\mu_{\text{old}}$. The optimal mapping P_t for our AGA admits a Sylvester-form [32] solution (Theorem 1), which penalizes transport mismatch along low-variance directions while providing anisotropic shrinkage to limit noise amplification.

Theorem 1 (AGA formulation via Ridge-GLS and anisotropic shrinkage). *Let $z_{\text{old}}, z_{\text{new}} \in \mathbb{R}^d$ denote paired features, and define centered variables $x = z_{\text{old}} - \mu_{\text{old}}$, $y = z_{\text{new}} - \mu_{\text{new}}$ with $\Sigma_{\text{old}} = \text{Cov}(x) \succ 0$, $\Sigma_{\text{new}} = \text{Cov}(y) \succ 0$, and $C_{\text{on}} = \text{Cov}(x,y)$. For $\rho > 0$, the ridge-regularized Mahalanobis regression (Eq. 5) leads to the unique minimizer satisfying $b^\star = \mu_{\text{new}} - P^\star\mu_{\text{old}}$, where $P^\star$ is the unique solution of the Sylvester equation:*

$$P^\star\Sigma_{\text{old}} + \rho\,\Sigma_{\text{new}}P^\star = C_{\text{on}}^\top. \tag{6}$$

Moreover, letting $\Sigma_{\text{new}} = U\Lambda U^\top$ and $\Sigma_{\text{old}} = V\Gamma V^\top$ with $\Lambda = \text{diag}(\lambda_i)$ and $\Gamma = \text{diag}(\gamma_j)$, the solution admits the spectral form $P^\star = UXV^\top$ where

$$X_{ij} = \frac{(U^\top C_{\text{on}}^\top V)_{ij}}{\gamma_j + \rho\,\lambda_i}. \tag{7}$$

Topology-Aware Evolution and the Overall Objective. While the AGA aligns the macroscopic geometry, directly superimposing it on an unconstrained, rapidly evolving backbone can introduce instability. Open-loop optimization of f_t naturally induces local topological degradation of the manifold, creating non-linear local distortions that the global linear prior $\bar{P}_t$ cannot capture.

To bridge this gap, we integrate a Topology-Aware Evolution objective directly into the active training phase. During training, the AGA prior used in this objective is the EMA-smoothed affine map $\bar{A}_t^{\text{lin}}(z) = \bar{P}_t z + \bar{b}_t$. We define a joint topological penalty $\mathcal{L}_{\text{top}}$ that couples the forward residual network g and the backward distiller D:

$$\mathcal{L}_{\text{top}} = \underbrace{\left\| g(z_{\text{old}};\theta_t) - \text{stopgrad}\left(z_{\text{new}} - \overbrace{\bar{A}_t^{\text{lin}}(z_{\text{old}})}^{\textbf{AGA Prior}} \right) \right\|_2^2}_{\text{forward residual calibration}} + \underbrace{\|D(z_{\text{new}}) - z_{\text{old}}\|_2^2}_{\text{backward manifold anchoring}} . \tag{8}$$

The overall end-to-end training objective for the active backbone and transport modules is thus formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{top}}\mathcal{L}_{\text{top}} + \mathcal{L}_{\text{ac}}, \tag{9}$$

where $\mathcal{L}_{\text{ac}}$ denotes the anti-collapse loss proposed by AdaGauss [29]. This unified formulation explicitly binds the network's evolution to the geometric prior. In $\mathcal{L}_{\text{top}}$, the forward residual calibration directs the shallow MLP $g(\cdot;\theta_t)$ to absorb the non-linear displacement that the analytic AGA does not capture. The stopgrad operator ensures that the residual tracks the evolving space without artificially limiting the backbone's plasticity for the new task.

Simultaneously, the backward manifold anchoring restricts structural collapse. Instead of strictly penalizing Euclidean drift (which limits plasticity), we employ the auxiliary distiller D to ensure the new representation retains sufficient geometric integrity to reconstruct the legacy manifold. Its gradients flow directly into f_t, penalizing the backbone for deformations that would disrupt legacy neighborhoods. Together, these joint constraints restrict the backbone to topology-preserving deformations, ensuring the representation respects the bounds required for stable Bayes evaluation (Theorem 2).

3.3 EMA-Updating the AGA During Backbone Training

Because f_t is optimized dynamically via SGD, the joint distribution of $(z_{\text{old}}, z_{\text{new}})$ evolves continuously. Instantaneous closed-form estimates of (P_t, b_t) can exhibit high-variance fluctuations, which may destabilize the residual calibration objective. We therefore maintain an exponential moving average (EMA) of the linear prior.

Periodic prior refresh & EMA smoothing. Every Δ iterations, we recompute an instantaneous GLS estimate $(\widehat{P}_t^{(k)}, \widehat{b}_t^{(k)})$ from a short stream of

paired features. We update a running prior $(\bar{P}_t, \bar{b}_t)$ for the AGA with momentum $m \in [0, 1)$:

$$\bar{P}_t \leftarrow m\,\bar{P}_t + (1-m)\,\widehat{P}_t^{(k)}, \qquad \bar{b}_t \leftarrow m\,\bar{b}_t + (1-m)\,\widehat{b}_t^{(k)}. \tag{10}$$

Between refreshes, the adapter uses $\bar{A}_t^{\mathrm{lin}}(z) = \bar{P}_t z + \bar{b}_t$. By treating $(\bar{P}_t, \bar{b}_t)$ as constants during backpropagation, the EMA prior provides a smooth, lower-variance geometric baseline.

3.4 Theoretical Guarantees for Bayesian Evaluation

As defined in Sec. 3.1, evaluation relies on Mahalanobis scores computed from class-conditional Gaussian statistics in the active feature space. The reliability of this evaluation depends on how accurately old-class moments are transported through A_t. We formalize this dependency by establishing a stability bound on the decision margin:

Theorem 2 (Margin stability under transported moment errors). *After task t, let the Bayes/Mahalanobis score be $d_c(z) = (z-\mu_c)^\top \Sigma_c^{-1}(z-\mu_c)$, with prediction $\hat{y}(z) = \arg\min_c d_c(z)$. Define the old-class margin for a labeled point (z, y) as $m(z) = \min_{c\neq y} d_c(z) - d_y(z)$. Assume for each class c we use transported parameters $(\hat{\mu}_c, \hat{\Sigma}_c)$ with perturbations $\Delta\mu_c = \hat{\mu}_c - \mu_c$ and $\Delta\Sigma_c = \hat{\Sigma}_c - \Sigma_c$ satisfying $\|\Delta\mu_c\|_2 \leq \eta_\mu$ and $\|\Delta\Sigma_c\|_2 \leq \eta_\Sigma$. Let $\lambda_{\min,c}$ denote the smallest eigenvalue of Σ_c and assume $\eta_\Sigma < \frac{1}{2}\lambda_{\min,c}$ for all c.*

Then for any feature z, the deviation of each class score is uniformly bounded by:

$$\begin{aligned} |\hat{d}_c(z) - d_c(z)| \leq & \underbrace{\frac{2}{\lambda_{\min,c}}\|z-\mu_c\|_2\,\eta_\mu + \frac{1}{\lambda_{\min,c}}\eta_\mu^2}_{\textit{mean error term}} \\ & + \underbrace{\frac{2\,\eta_\Sigma}{\lambda_{\min,c}^2}(\|z-\mu_c\|_2 + \eta_\mu)^2}_{\textit{covariance error term}}. \end{aligned} \tag{11}$$

Consequently, letting $\varepsilon_c(z)$ denote the bound above and $\varepsilon(z) = \varepsilon_y(z) + \max_{c\neq y}\varepsilon_c(z)$, we guarantee:

$$m(z) > \varepsilon(z) \quad \Longrightarrow \quad \arg\min_c \hat{d}_c(z) = y, \tag{12}$$

meaning the Bayes/Mahalanobis prediction remains identical despite the transport perturbations.

Implications for our framework. Theorem 2 connects robust Mahalanobis evaluation to the accuracy of transported means and covariances. The AGA suppresses macroscopic anisotropic drift through the Mahalanobis-aligned closed-form prior (Theorem 1), while the topology-aware objective limits local manifold

degradation during backbone optimization. Together, these two components reduce the perturbation terms η_μ and η_Σ that directly control the margin-stability condition.

After task training, the final adapter is instantiated as

$$A_t(z) = \bar{P}_t z + \bar{b}_t + g(z; \theta_t). \tag{13}$$

The analytic term accounts for the dominant global transport, and the residual term captures the remaining localized deformation. As defined in Sec. 3.1, old Gaussian statistics are then pushed forward through A_t in a training-free stage. This design keeps transport coupled with representation learning during optimization, while avoiding an additional post-hoc adapter fine-tuning stage. Our ablation in Sec. 4.2 further verifies that the end-to-end adapter is already effective after co-evolution, and that optional post-hoc refinement is generally unnecessary.

4 Experiments

Evaluation setup and baselines. Following standard exemplar-free protocols, we compare against regularizers (EWC [15], LwF [21]), prototype approaches (SDC [33], PASS [37], FeTrIL [28], FeCAM [6]), and recent methods (EFC [23], ADC [7], LDC [5], AdaGauss [29], DPCR [11]). Baselines are reproduced using PyCIL [36], FACIL [25], OCL [24], or their official codes. We evaluate on CIFAR-100 [18], TinyImageNet [20], ImageNet-100 [3] (from scratch) and CUB-200 [31] (pretrained) over $T \in \{10, 20\}$ tasks, reporting A_{last} and running average A_{inc}.

Implementation. Our codebase builds upon AdaGauss [29], preserving its prototype/Gaussian bookkeeping and classifier to isolate performance improvements to the proposed transport mechanism. Unless stated otherwise, we train a ResNet-18 [10] with batch size 256, project 512-D features into a 64-D prototype space, and maintain class-conditional Gaussians with a Gaussian Bayes classifier. Optimization schedules follow AdaGauss (SGD, 200 epochs for CIFAR-100/TinyImageNet/ImageNet-100; CUB-200 uses a pretrained backbone with a separate head learning rate). Detailed hyperparameter settings and sensitivity experiments can be found in the Supplementary Material. We additionally verify in the Supplementary Material that a sampling-based linear classifier trained from the same transported Gaussians gives comparable performance, indicating that the improvement comes from better transported statistics rather than a specific readout.

Reproduction note. All reproduced baselines are evaluated under the same class orders and protocol. During reproduction of covariance- or projection-based methods, we observed that some public implementations can encounter rank-deficient or nearly singular covariance/projection matrices, which may break SVD, inverse, pseudo-inverse, or Gaussian construction routines. We therefore apply standard numerical safeguards such as symmetrization, diagonal loading,

Table 1: CIFAR-100 & TinyImageNet (from scratch). Average incremental (A_{inc}) and last-task average (A_{last}) accuracy (%). Results are averaged over five runs (standard deviations are provided in Supplementary Material). Best in **bold**.

Method	CIFAR-100				TinyImageNet			
	T=10		T=20		T=10		T=20	
	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}
EWC	30.9	50.4	17.0	34.2	18.5	34.3	11.3	26.8
LwF[ECCV16]	31.9	51.8	17.6	39.2	27.1	39.6	15.2	31.5
SDC[CVPR20]	40.6	56.2	32.3	46.6	29.5	43.8	26.3	40.6
PASS[CVPR21]	30.8	48.3	17.6	31.1	24.5	39.5	18.5	30.4
FeTrIL[WACV23]	34.9	51.2	23.3	37.9	31.0	45.3	25.9	39.9
FeCAM[NeurIPS23]	32.4	48.7	21.1	34.5	30.9	44.9	24.9	37.9
EFC[ICLR24]	43.5	58.1	32.4	47.0	34.5	47.9	28.4	42.1
ADC[CVPR24]	46.5	61.4	35.1	51.7	32.3	43.0	18.1	36.0
LDC[ECCV24]	45.4	59.5	35.5	51.9	34.2	46.8	24.9	38.2
AdaGauss[NeurIPS24]	46.8	60.9	37.9	54.4	32.9	45.8	27.5	39.5
DPCR[ICML2025]	50.2	62.8	39.8	54.8	34.3	46.9	25.6	39.3
Ours	**52.0**	**64.4**	**43.0**	**56.6**	**37.2**	**50.6**	**31.5**	**45.1**

eigenvalue clipping, and jitter when needed. These safeguards preserve the original objectives and are applied only to ensure stable execution. For CUB-200, DPCR did not complete reliably under the pretrained protocol, so we leave these entries blank rather than reporting selectively successful runs.

4.1 Main Results

Tables 1 and 2 present the mean accuracy of our framework (Ours) alongside established baselines. Standard deviations across five runs are provided in Supplementary Material. Across the from-scratch benchmarks, our method yields consistent improvements over the evaluated competitors.

On **CIFAR-100**, our method achieves competitive performance across both settings. At T=10, it improves upon DPCR by +1.8 pp in A_{last} and +1.6 pp in A_{inc}, and exceeds AdaGauss by +5.2 pp and +3.5 pp, respectively. A similar trend of final-task retention is observed at T=20, where it leads DPCR by +3.2 pp in A_{last}.

On **TinyImageNet**, similar performance gains are observed. Our method improves upon DPCR by +2.9/+3.7 pp at T=10 (for $A_{\text{last}}/A_{\text{inc}}$). At T=20, it exceeds EFC by +3.1/+3.0 pp, suggesting that the geometry-aware transport is effective at managing representation drift over longer task sequences.

On **ImageNet-100** (from scratch), our framework maintains consistent performance. It achieves the highest final-task accuracy (A_{last}) at both T=10 (53.2%) and T=20 (44.7%), outperforming baselines like AdaGauss and EFC. While LDC

Table 2: ImageNet-100 (from scratch) & CUB-200 (pretrained). Average incremental (A_{inc}) and last-task average (A_{last}) accuracy (%). †: results reported by [5]. Results are averaged over five runs (standard deviations are provided in Supplementary Material). Best in **bold**.

Method	ImageNet-100				CUB-200			
	T=10		T=20		T=10		T=20	
	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}
EWC	25.1	40.6	13.7	29.2	15.8	32.6	12.3	27.2
LwF[ECCV16]	33.4	51.5	18.6	41.3	30.4	46.1	19.4	34.7
SDC[CVPR20]	35.4	50.1	19.4	36.5	50.3	60.5	27.9	40.1
PASS[CVPR21]	26.4	45.7	14.4	31.7	27.0	42.3	18.1	36.9
FeTrIL[WACV23]	36.2	52.6	26.6	42.4	36.9	48.2	34.6	45.3
FeCAM[NeurIPS23]	38.7	54.8	29.0	44.6	40.2	54.9	36.2	48.9
EFC[ICLR24]	50.9	61.3	38.6	50.5	51.0	63.3	**46.1**	**59.3**
ADC[CVPR24]	38.3	55.5	25.1	43.4	49.5	58.8	35.4	48.3
LDC[ECCV24]	51.4†	**69.4**†	28.5	46.5	47.5	55.7	27.2	39.8
AdaGauss[NeurIPS24]	51.1	65.0	42.6	57.4	**52.9**	**63.4**	45.0	57.0
DPCR[ICML2025]	49.9	64.8	37.3	54.7	–	–	–	–
Ours	**53.2**	67.4	**44.7**	**59.1**	52.4	63.1	43.7	55.8

reports a higher running average (A_{inc}) at T=10, our approach remains competitive incrementally while demonstrating higher final knowledge retention.

Finally, on **CUB-200** with a pretrained backbone, representation drift is typically milder, leaving less headroom for cross-task alignment. Our method performs comparably to AdaGauss at T=10, while EFC demonstrates the strongest performance at T=20 (46.1/59.3). Note that DPCR results are omitted (marked as "–") because it did not complete reliably under the pretrained CUB-200 protocol, as discussed in the reproduction note above. A detailed discussion of the overall CUB-200 results and the effects of pretrained backbones is provided in the Supplementary Material. Overall, these evaluations validate the efficacy of the proposed framework in managing representation drift across diverse incremental learning scenarios.

4.2 Component Study: Validating the Geometry-Anchored Framework

Our framework integrates an Analytic Geometric Anchor (AGA) into a Topology-Aware Evolution process. Since the AGA operates as an explicit closed-form prior, it relies on the topology-aware training phase to function effectively. We sequentially ablate these components in Table 3 to evaluate their respective contributions.

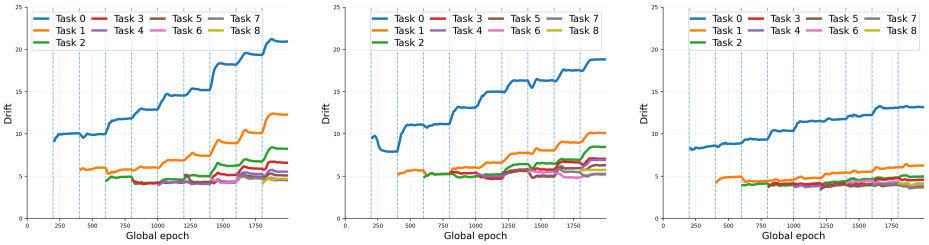
Table 3: Component ablation of the Geometry-Anchored Transport Framework on CIFAR-100 and TinyImageNet (T=10 and T=20). We ablate the Topology-Aware Evolution (**Top-Aware**), the Analytic Geometric Anchor (**AGA**), and the optional post-hoc refinement (**Refine**). Best results in **bold**.

Components			**CIFAR-100**				**TinyImageNet**			
			T=10		T=20		T=10		T=20	
Top-Aware	AGA	Refine	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}	A_{last}	A_{inc}
×	×	✓	46.8	60.9	37.9	54.4	32.9	45.8	27.5	39.5
✓	×	✓	50.3	63.1	41.6	55.9	35.6	49.3	30.4	44.2
✓	✓	✓	51.7	64.3	42.3	56.3	37.0	50.4	**31.7**	**45.3**
✓	✓	×	**52.0**	**64.4**	**43.0**	**56.6**	**37.2**	**50.6**	31.5	45.1

Impact of the Analytic Geometric Anchor. The baseline (× Top-Aware, × AGA, ✓ Refine) structurally aligns with the AdaGauss paradigm: open-loop optimization followed by a decoupled adapter. Upgrading this to include topology-aware constraints (✓ Top-Aware, × AGA, ✓ Refine) allows the adapter to be learned during the primary training phase. However, without the AGA prior, the adapter effectively functions as a standard MLP mapping $z_{\text{old}} \to z_{\text{new}}$. While this restricts manifold distortion—improving CIFAR-100 A_{last} from 46.8% to 50.3%—relying solely on this neural projection is limited. Such a mapping optimizes a standard Euclidean objective, which may not fully account for the Mahalanobis-sensitive tail errors that impact Bayes evaluation.

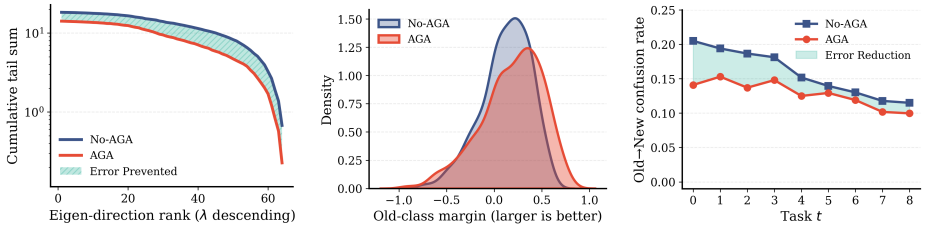
Comparing end-to-end evolution and post-hoc refinement. Performance improves further when the AGA is embedded as a geometric prior, transforming the adapter's role from a direct projection to a localized residual correction. Notably, our results suggest that decoupled post-hoc fine-tuning does not necessarily improve well-calibrated representations. As shown in Table 3, the end-to-end variant without secondary refinement (gray row: × Refine) generally performs better than the version with refinement (✓ Refine) across most settings, reaching 52.0% on CIFAR-100.

This indicates that the geometry-anchored framework reaches a stable configuration during the active training phase. Isolated fine-tuning of the adapter after freezing the backbone may disrupt the coupling established during co-evolution, potentially causing the adapter to overfit to static features. An exception is observed in the longer task sequence of TinyImageNet (T=20), where accumulated drift makes isolated calibration marginally beneficial (31.7% vs 31.5%). Overall, these results indicate that embedding transport as a primary training constraint provides a robust one-stage framework, reducing the need for retroactive fine-tuning.



(a) AdaGauss **(b)** Ours (w/o AGA) **(c)** Ours (Full)

Fig. 3: Prototype drift over global epochs on Tiny-ImageNet (10 steps). We track, for each past task k, the mean feature $\mu_c^{(e)}$ of class c under the current backbone at global epoch e against its stored/transported prototype $\hat{\mu}_c^{(e)}$ used for Bayes evaluation, reporting $\text{Drift}_k(e) = \frac{1}{|\mathcal{C}_k|}\sum_{c\in\mathcal{C}_k}\|\mu_c^{(e)} - \hat{\mu}_c^{(e)}\|_2$. The topology-aware anchoring alone (b) reduces drift relative to AdaGauss (a); calibrating the residual to the explicit AGA prior (c) further suppresses drift, particularly limiting error accumulation for earlier tasks.



(a) Maha-tail (Task 0) **(b)** Margin density (Task 0) **(c)** Old→New rate

Fig. 4: Diagnostic ablations for the Analytic Geometric Anchor (AGA). (a) Cumulative Mahalanobis tail T_t; the shaded region highlights the tail error mitigated by the AGA. **(b)** Estimated probability density of old-class margins; integrating the AGA shifts the distribution toward positive values. **(c)** Old→New confusion rate ρ_t across tasks; the geometric prior reduces misclassifications of legacy samples. Together, these diagnostics suggest that targeting Mahalanobis-sensitive tail error translates into more stable decision boundaries.

Our framework reduces prototype drift. We first evaluate the preservation of stored class prototypes. Fig. 3 tracks prototype drift over global epochs by comparing the current class-mean feature to the corresponding transported prototype used for Bayes evaluation. While the anchored baseline variant (Ours w/o AGA) limits initial topological changes compared to AdaGauss, enforcing the explicit AGA prior (Ours) further suppresses this drift. This stabilization is noticeable for earlier tasks, where representation changes accumulate over time, indicating that the analytic anchor helps maintain faithfulness in the old→new transport.

Suppressing Mahalanobis-sensitive tail error (Fig. 4a). We report the cumulative Mahalanobis tail T_t, which characterizes the drift sensitivity across the feature spectrum (formulated in the Supplementary Material). The Ma-

halanobis metric assigns larger weights to feature directions with naturally small variance (the spectrum's "tail"). Consequently, parameter drifts along these directions are amplified, contributing to forgetting. The transport curve for our full framework (Ours) bounds the non-AGA baseline (Ours w/o AGA) from below. The shaded region reveals that the magnitude of error mitigated by the AGA increases at high eigen-ranks (the tail where $\lambda_i \to 0$). This observation supports the anisotropic shrinkage established in Theorem 1: governed by the spectral denominator $\gamma_j + \rho\,\lambda_i$, the closed-form prior dampens updates along sensitive dimensions, mitigating the error components that affect Mahalanobis evaluation.

Preserving old-class decision margins (Fig. 4b). We estimate the old-class margin $m_t(x)$, representing the log-likelihood gap between the ground-truth and the nearest competing class (refer to the SUPPLEMENTARY for the precise definition). A larger positive margin indicates higher robustness against new-class interference. As shown, the geometry-anchored framework shifts the density peak toward larger, positive values compared to the baseline. This separation between the legacy class and competitors aligns with the tail-error mitigation observed in Fig. 4a. These findings support our theoretical premise (Theorem 2): targeting Mahalanobis-sensitive tail error helps stabilize decision boundaries.

Mitigating old→new confusions (Fig. 4c). We track the Old→New confusion rate ρ_t, which measures the empirical probability of misclassifying legacy samples into newly introduced classes (detailed in the SUPPLEMENTARY MATERIAL). The addition of the AGA consistently reduces ρ_t across the continual learning trajectory relative to the non-AGA baseline. The shaded area illustrates the framework's capability in limiting legacy samples from being misclassified into new classes. These improvements are consistent with the tail suppression and margin preservation established in earlier diagnostics.

Overhead of online prior estimation. Our framework periodically refreshes the transport prior during training. For E training epochs per task and a refresh interval of K epochs, we perform $R \approx E/K$ refreshes. Each refresh samples B_t mini-batches, requiring two additional forward passes per batch (one each for the frozen and current backbones). To maintain consistent relative compute across varying task partitions, we scale B_t proportionally to the optimization steps per epoch S_t (i.e., $B_t = \gamma S_t$). The relative computational overhead therefore scales as:

$$\frac{2RB_t}{cES_t} \;\approx\; \frac{2\gamma}{cK},$$

rendering it relatively independent of the specific task partition. Under our default configuration (E=200, K=10), and taking one refresh to sample approximately one epoch's worth of steps (γ=1) with a standard approximation that one training step costs about three forward passes (c≈3), the relative overhead is:

$$\frac{2\gamma}{cK} = \frac{2}{3 \cdot 10} = \frac{1}{15} \approx 6.7\%.$$

This corresponds to a moderate single-digit percentage overhead. Furthermore, because the GLS closed-form solution (Theorem 1) operates exclusively on low-dimensional feature covariances rather than high-dimensional image tensors, the matrix solver overhead is virtually negligible. These factors confirm that analytic anchoring remains highly computationally efficient in practice.

5 Conclusion, Limitations, and Future Work

Conclusion. We introduced the Geometry-Anchored Transport Framework, formulating feature transport in exemplar-free class-incremental learning as an integrated constraint during representation learning rather than a strictly post-hoc procedure. Our method combines an Analytic Geometric Anchor with Topology-Aware Evolution to regularize the legacy manifold, while a residual network calibrates the non-linear deformations. This coupled design facilitates the transport of cached Gaussian statistics for classification, addressing the sensitivity of decision margins to moment-transport errors.

Limitations. Our approach is currently designed for prototype preservation with Bayes-style evaluation, and its theoretical guarantees are aligned with this parametric setting. The analytic anchor relies on an affine prior with EMA updates; while the residual network accommodates local non-linearities, severe non-linear or highly non-stationary representation shifts may still challenge this linearized global approximation, particularly when covariance estimates are noisy or poorly conditioned.

Future work. Future research could explore generalizing geometry-anchored transport beyond single class-conditional Gaussians (e.g., using Gaussian mixtures [17] or implicit density surrogates [35]) while maintaining margin stability bounds. Additionally, investigating alternative anchor formulations (e.g., low-rank, kernelized, or orthogonality-constrained priors [16]) may offer ways to model complex global deformations while preserving the separation of analytic priors and neural residuals.

Supplementary Material. Due to space constraints, extended details are deferred to the supplementary material. Key contents include: (1) complete mathematical proofs for Theorems 1 and 2; (2) a unified formulation of prior post-hoc feature transport methods; (3) algorithmic pseudocode, architectural specifications (e.g., feature dimensions), and an extended analysis of MLP computational overhead; and (4) comprehensive experimental configurations, encompassing full hyperparameter settings, sensitivity analyses, formal diagnostic metric definitions, and our rationale for baseline (AdaGauss) selection. Additionally, the supplementary material reports classifier-robustness results with an alternative sampling-based linear head, showing that the gains persist beyond the default Mahalanobis readout.

References

1. Boschini, M., Bonicelli, L., Buzzega, P., Porrello, A., Calderara, S.: Class-incremental continual learning into the extended der-verse. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5497–5512 (2023)
2. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: ECCV (2018)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
4. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: ECCV (2020)
5. Gomez-Villa, A., Goswami, D., Wang, K., Andrew, B., Twardowski, B., van de Weijer, J.: Exemplar-free continual representation learning via learnable drift compensation. In: ECCV (2024)
6. Goswami, D., Liu, Y., Twardowski, B., van de Weijer, J.: Fecam: Feature covariance and mahalanobis metric for incremental learning. In: NeurIPS (2023)
7. Goswami, D., Soutif-Cormerais, A., Liu, Y., Kamath, S., Twardowski, B., van de Weijer, J.: Resurrecting old classes with new data for exemplarfree continual learning. In: CVPR (2024)
8. Hacohen, G., Tuytelaars, T.: Predicting the susceptibility of examples to catastrophic forgetting. In: ICML (2025)
9. Hadsell, R., Rao, D., Rusu, A.A., Pascanu, R.: Embracing change: Continual learning in deep neural networks. Trends in Cognitive Sciences (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. He, R., Fang, D., Xu, Y., Cui, Y., Li, M., Chen, C., Zeng, Z., Zhuang, H.: Semantic shift estimation via dual-projection and classifier reconstruction for exemplar-free class-incremental learning. In: ICML (2025)
12. Kanebako, Y., Tsukamoto, K.: Puzzleshuffle: Undesirable feature learning for semantic shift detection. In: Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part V. Lecture Notes in Computer Science, vol. 12979, pp. 3–19. Springer (2021)
13. Kim, D., Han, B.: On the stability-plasticity dilemma of class-incremental learning. In: CVPR (2023)
14. Kim, J., Kim, Y., Sohn, J.: Measuring representational shifts in continual learning: A linear transformation perspective. In: ICML (2025)
15. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al.: Overcoming catastrophic forgetting in neural networks. In: Proceedings of the National Academy of Sciences (2017)
16. Koh, H., Seo, M., Bang, J., Song, H., Hong, D., Park, S., Ha, J., Choi, J.: Online boundary-free continual learning by scheduled data prior. In: ICLR (2023)
17. Korycki, L., Krawczyk, B.: Class-incremental mixture of gaussians for deep continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2024)
18. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
19. Lange, M.D., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G.G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3366–3385 (2022)

20. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS231N (2015)
21. Li, Z., Hoiem, D.: Learning without forgetting. In: ECCV (2016)
22. Lyu, Y., Wang, L., Zhang, X., Sun, Z., Su, H., Zhu, J., Jing, L.: Overcoming recency bias of normalization statistics in continual learning: Balance and adaptation. In: NeurIPS (2023)
23. Magistri, S., Trinci, T., Soutif-Cormerais, A., van de Weijer, J., Bagdanov, A.D.: Elastic feature consolidation for cold start exemplarfree incremental learning. In: ICLR (2024)
24. Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., Sanner, S.: Online continual learning in image classification: An empirical survey. Neurocomputing **469**, 28–51 (2022)
25. Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A.D., van de Weijer, J.: Class-incremental learning: Survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 5513–5533 (2023)
26. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of learning and motivation **24**, 109–165 (1989)
27. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019)
28. Petit, G., Popescu, A., Schindler, H., Picard, D., Delezoide, B.: Fetril: Feature translation for exemplar-free class-incremental learning. In: WACV (January 2023)
29. Rypeść, G., Cygert, S., Trzciński, T., Twardowski, B.: Task-recency bias strikes back: Adapting covariances in exemplar-free class incremental learning. In: NeurIPS (2024)
30. Serra, J., Suris, D., Miron, M., Karatzoglou, A.: Overcoming catastrophic forgetting with hard attention to the task. In: ICML (2018)
31. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
32. Wu, A., Feng, G., Duan, G., Wu, W.: Closed-form solutions to sylvester-conjugate matrix equations. Comput. Math. Appl. **60**(1), 95–111 (2010)
33. Yu, L., Twardowski, B., Liu, X., Herranz, L., Wang, K., Cheng, Y., Jui, S., van de Weijer, J.: Semantic drift compensation for class-incremental learning. In: CVPR (2020)
34. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: ICML (2017)
35. Zhang, Y., Zhang, Z., Zhao, P., Sugiyama, M.: Adapting to continuous covariate shift via online density ratio estimation. In: NeurIPS (2023)
36. Zhou, D.W., Wang, F.Y., Ye, H.J., Zhan, D.C.: Pycil: a python toolbox for class-incremental learning. SCIENCE CHINA Information Sciences **66**(9), 197101 (2023)
37. Zhu, F., Zhang, X.Y., Wang, C., Yin, F., Liu, C.L.: Prototype augmentation and self-supervision for incremental learning. In: CVPR (2021)
38. Zhuang, H., He, R., Tong, K., Zeng, Z., Chen, C., Lin, Z.: Ds-al: A dual-stream analytic learning for exemplar-free class-incremental learning. In: AAAI (2024)
39. Zhuang, H., Weng, Z., Wei, H., Xie, R., Toh, K.A., Lin, Z.: Acil: Analytic class-incremental learning with absolute memorization and privacy protection. In: NeurIPS (2022)