DYNAMIC SEMANTIC EVOLUTION FOR SEMI-SUPERVISED FEW-SHOT INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-Supervised Few-Shot Class Incremental Learning (SSFSCIL) aims to continuously acquire novel concepts from scarce annotations and abundant unlabeled data while preserving previously learned knowledge. However, existing approaches mainly rely on static pseudo-labeling and incremental adaptation, which struggle with three fundamental challenges: unreliable semantic discovery under limited supervision, unstable representation evolution across sessions, and insufficient decision boundary characterization with few-shot samples. In this paper, we introduce DSE-SSF, a novel framework that reformulates continuous SSFSCIL as a process of dynamic semantic evolution rather than sequential parameter optimization. Based on this insight, DSE-SSF establishes a closed-loop learning paradigm consisting of semantic discovery, knowledge stabilization, and boundary refinement, implemented by three novel mechanisms: (1) Dynamic Semantic Graph Evolution progressively uncovers reliable semantic structures from unlabeled data through evolving manifold propagation; (2) Adaptive Dual-Expert Stabilization leverages temporal consistency and adaptive uncertainty calibration to prevent incremental knowledge drift; and (3) Boundary-aware Feature Manifold Exploration explicitly investigates uncertain feature regions to construct robust decision boundaries under sparse supervision. By jointly modeling the evolution of semantic topology, representation stability, and boundary geometry, DSE-SSF enables reliable knowledge acquisition throughout continuous incremental learning. Extensive experiments on SSFSCIL benchmarks demonstrate that DSE-SSF consistently outperforms existing methods, validating the effectiveness of dynamic semantic evolution for long-term few-shot visual learning.

1 INTRODUCTION

Deep neural networks have achieved remarkable success in visual recognition, yet their ability to continuously acquire new knowledge remains fundamentally limited. In real-world scenarios, visual concepts emerge progressively, where new categories appear over time while previously learned knowledge must be preserved Zenke et al. (2017); Parisi et al. (2019). This setting gives rise to class incremental learning (CIL), which aims to continuously expand recognition capability without accessing historical data Jodelet et al. (2023). However, conventional CIL methods typically assume sufficient labeled samples for newly arriving classes, an assumption that rarely holds in practical applications where annotation costs are prohibitively expensive Achille et al. (2018).

Few-Shot Class Incremental Learning (FSCIL) has recently emerged to address this challenge by restricting novel categories to only a few labeled examples Tao et al. (2020). Nevertheless, FSCIL remains highly dependent on supervised signals and suffers from severe representation degradation when encountering continuously emerging concepts. To alleviate annotation limitations, Semi-Supervised Few-Shot Class Incremental Learning (SSFSCIL) introduces abundant unlabeled data during each incremental session, providing an opportunity to exploit hidden semantic structures beyond scarce annotations Han et al. (2024). Despite its potential, existing SSFSCIL approaches largely inherit the paradigm of conventional semi-supervised learning, where unlabeled samples are primarily treated as static auxiliary signals through pseudo-label generation and consistency regularization Han et al. (2024). Such approaches overlook a fundamental characteristic of continual learning in which the semantic structure of the environment is not fixed, but continuously evolves as new concepts are introduced and representations are updated Jodelet et al. (2023). Therefore, SSFSCIL should not be viewed as a simple combination of semi-supervised learning and incremental

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

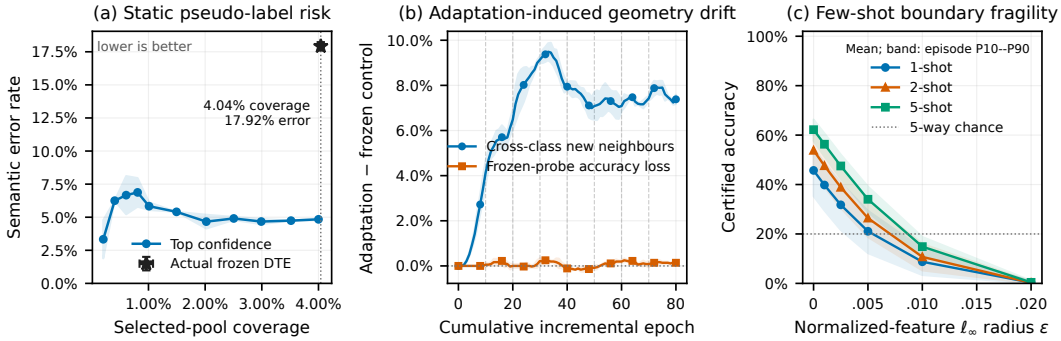

Figure 1: Motivation diagnostics on the reproduced SSFSCIL-DTE baseline. (a) Static pseudo-label risk, (b) adaptation-induced representation drift, and (c) few-shot boundary fragility. Hidden labels are used only for post-hoc evaluation. Curves report the mean over three seeds; bands show seed standard deviation, except the episode P10–P90 bands in (c). The complete experimental settings and extended diagnostic results are provided in Appendix C.4.

learning. Instead, it represents a fundamentally different learning paradigm, where a model must continuously discover, stabilize, and refine evolving semantic knowledge under extremely limited supervision.

We investigate the model's optimization behavior and empirically observe that optimizing a model in SSFSCIL faces three challenges: (1) As shown in Fig. 1(a), static confidence-based pseudo-labeling still introduces considerable semantic errors, indicating that prediction confidence alone is insufficient for reliable unlabeled data exploitation; (2) The results of Fig. 1(b) show that incremental adaptation causes continuous representation drift, leading to instability between old and new knowledge; (3) The results of Fig. 1(c) show that few-shot supervision results in fragile decision boundaries, where robustness rapidly decreases under feature perturbations. These three challenges interact with each other, preventing the model from achieving the best performance.

However, existing methods typically address only individual aspects of these problems under SSFSCIL Han et al. (2024), as shown in Fig. 4, lacking a unified mechanism that models the evolution of semantic knowledge throughout incremental learning. To bridge this gap, we propose Dynamic Semantic Evolution for Semi-Supervised Few-Shot Incremental Learning (DSE-SSF), a new framework that reformulates SSFSCIL as a process of dynamic semantic evolution rather than sequential parameter adaptation. The key insight of DSE-SSF is that reliable incremental learning requires a closed-loop evolutionary process consisting of semantic discovery, knowledge stabilization, and decision boundary refinement, each of which addresses one specific challenge.

Based on this principle, DSE-SSF introduces three complementary components. (1) Dynamic Semantic Graph Evolution (DSGE) models unlabeled samples as evolving semantic nodes instead of isolated predictions. By continuously updating the semantic topology according to the evolving feature manifold, DSGE progressively discovers reliable semantic supervision beyond model confidence. (2) Adaptive Dual-Expert Stabilization (ADES) maintains two temporal perspectives of knowledge evolution, in which an active expert that adapts to novel concepts and a stabilized expert that preserves the historical semantic trajectory. Their discrepancy provides an adaptive uncertainty estimation mechanism, enabling reliable pseudo supervision selection and preventing unstable representation drift. (3) Boundary-aware Feature Manifold Exploration (BFME) explicitly explores uncertain regions around decision boundaries in the feature space. Instead of generating arbitrary perturbations, BFME searches informative semantic neighborhoods and enforces local consistency, improving classifier robustness under few-shot supervision.

Together, these components establish a dynamic semantic evolution paradigm, where semantic structures, representations, and decision boundaries are jointly optimized throughout incremental learning. Extensive experiments on SSFSCIL benchmarks demonstrate that DSE-SSF consistently surpasses existing approaches, validating the effectiveness of modeling incremental learning as continuous semantic evolution.

Our contributions are summarized as follows: (1) We introduce a new perspective for continuous SSFSCIL by formulating incremental learning as an evolving semantic knowledge acquisition process, rather than merely sequential parameter optimization; (2) We propose DSE-SSF, a unified

framework consisting of Dynamic Semantic Graph Evolution, Adaptive Dual-Expert Stabilization, and Boundary-aware Feature Manifold Exploration, which jointly address semantic ambiguity, incremental instability, and boundary deficiency; (3) We develop a dynamic semantic discovery mechanism that exploits the evolving topology of unlabeled data and an uncertainty-aware stabilization strategy that adaptively regulates pseudo supervision during incremental adaptation.

2 Related Work

Few-Shot Class Incremental Learning extends class-incremental learning to data-scarce novel sessions, where overfitting and forgetting must be controlled jointly. TOPIC organizes embeddings with a neural gas, while CEC evolves classifiers through graph attention Tao et al. (2020); Zhang et al. (2021b). A complementary line prepares the representation space for future classes: C-FSCIL aligns quasi-orthogonal prototypes, whereas FACT and SAVC reserve capacity using virtual prototypes or classes Hersche et al. (2022); Zhou et al. (2022); Song et al. (2023). CABD transfers knowledge from complementary teachers; TEEN calibrates novel prototypes without retraining; and DPCL couples dynamic pseudo-label assignment with prototypical learning Zhao et al. (2023); Wang et al. (2023); Mao et al. (2026). These methods strengthen supervised FSCIL, but largely treat the few labeled exemplars as the only novel-session evidence; DSE-SSF instead targets the abundant unlabeled stream.

Semi-Supervised Few-Shot Class Incremental Learning introduces an unlabeled set in every novel session. Early SSFSCIL combines self-training with knowledge distillation, and Us-KD and UaD-ClE improve it through uncertainty-guided selection and adaptive distillation Cui et al. (2021; 2023; 2024). DTE moves beyond independent predictions by propagating labels over Voronoi partitions and Hebbian category graphs Han et al. (2024). In the broader semi-supervised continual setting, DSGD distills dynamic subgraphs and PsHD preserves topological structure Fan et al. (2024b;a). These studies motivate exploiting unlabeled structure under incremental changes. DSE-SSF couples semantic discovery with temporal stabilization and boundary refinement throughout each incremental session.

Graph-Based Semi-Supervision and Consistency Learning exploits manifold smoothness and neighborhood-based label propagation Zhou et al. (2003); Iscen et al. (2019). Modern SSL combines label guessing, augmentation, and confidence control: MixMatch unifies label guessing with MixUp, Mean Teacher averages model weights, FixMatch couples thresholded pseudo labels with consistency, and FlexMatch adapts thresholds to learning progress Berthelot et al. (2019); Tarvainen & Valpola (2017); Sohn et al. (2020); Zhang et al. (2021a). VAT further encourages local smoothness under adversarial perturbations Miyato et al. (2019). Unlike these stationary single-model formulations, DSE-SSF jointly adapts graph structure, reliability estimation, and boundary regularization to evolving expert representations.

3 Methodology

3.1 Problem Formulation and Framework Overview

We investigate the problem of SSFSCIL, where a model is required to continuously acquire novel concepts from extremely limited labeled samples and abundant unlabeled data, while preserving previously learned knowledge. Following Han et al. (2024), the learning process consists of one base session and T incremental sessions. The base session contains a set of labeled classes $\mathcal{C}^0$ with sufficient annotations, while each incremental session t introduces a novel category set $\mathcal{C}^t$ with only K labeled samples per class and a large number of unlabeled samples. Specifically, the labeled and unlabeled data in session t are denoted as $\mathcal{D}_l^t$ and $\mathcal{D}_u^t$, respectively. The goal of SSFSCIL is to learn a unified classifier over all observed categories $\mathcal{C}_{all} = \bigcup_{t=0}^{T} \mathcal{C}^t$ without accessing data from previous sessions. Despite recent advances, SSFSCIL remains challenging due to three fundamental issues: (1) **semantic ambiguity**, where limited annotations make pseudo-label generation unreliable; (2) **incremental instability**, where continuously adapting the model causes representation drift; (3) **boundary deficiency**, where few-shot samples provide insufficient information to characterize class decision boundaries.

Overview of DSE-SSF. The overall framework of DSE-SSF is built upon a frozen visual foundation model with dynamically expandable lightweight experts, as shown in Fig. 2. Instead of continuously updating the backbone, we introduce task-specific LoRA experts, where each expert specializes in adapting the pretrained representation space toward newly emerging categories. Given an input image $\mathbf{x}$, the frozen backbone extracts a general visual representation $\mathbf{z} = f_\theta(\mathbf{x})$, where θ represents the frozen parameters.

For each incremental task t, a new expert $\mathcal{E}_t$ is dynamically created with lightweight adaptation parameters (the LoRA module). The learning process follows a progressive paradigm: (i) **Semantic Discovery:** Dynamic semantic graph evolution explores the manifold structure of unlabeled samples and generates reliable pseudo supervision. (ii) **Knowledge Stabilization:** Dual expert consistency learning constrains representation evolution and prevents incremental drift. (iii) **Boundary Refinement:** Boundary-aware feature manifold refinement explicitly explores uncertain regions around decision boundaries to improve generalization under few-shot conditions. Together, these components form a closed-loop optimization framework, where improved representations further refine semantic discovery, and reliable supervision continuously strengthens the incremental expert.

3.2 Dynamic Semantic Graph Evolution

In SSFSCIL, the limited availability of labeled samples makes conventional supervised optimization insufficient for learning novel concepts. Although unlabeled data provide abundant semantic information, existing pseudo-labeling strategies mainly rely on prediction confidence from the current model, which is unreliable under few-shot supervision and often suffers from confirmation bias.

To address this limitation, we propose a Dynamic Semantic Graph Evolution (DSGE) mechanism, which regards unlabeled samples as evolving semantic nodes rather than isolated predictions. Instead of directly assigning pseudo labels based on model confidence, DSGE progressively discovers reliable semantic structures by jointly exploiting model predictions and the intrinsic manifold relationship among samples. The key idea is that the semantic topology of unlabeled data should evolve together with the incremental expert. As the expert gradually improves its representation capability, the corresponding feature manifold becomes more discriminative, allowing the graph structure and pseudo supervision to be iteratively refined.

Semantic Graph Initialization. For an incremental session t, after the initial optimization of expert $\mathcal{E}_t$ using the limited labeled set $\mathcal{D}_l^t$, we combine the labeled and unlabeled samples $\mathcal{X}^t = \mathcal{D}_l^t \cup \mathcal{D}_u^t$. Each sample is regarded as a node in a semantic graph $G^t = (V^t, E^t)$, where the feature representation is extracted from the current expert $\mathbf{h}_i^t = F_{\text{map}}(\mathcal{E}_t, \mathbf{x}_i)$. The pairwise semantic affinity between two samples is measured by cosine similarity:

$$S_{ij}^t = \frac{(\mathbf{h}_i^t)^\top \mathbf{h}_j^t}{||\mathbf{h}_i^t||_2 ||\mathbf{h}_j^t||_2}, \tag{1}$$

where $||\cdot||_2$ denotes the L2 norm. Since few-shot incremental learning suffers from noisy feature relationships, we construct a sparse local semantic graph by retaining only the nearest neighbors:

$$A_{ij}^t = \begin{cases} S_{ij}^t, & j \in \mathcal{N}_K(i), \\ 0, & otherwise. \end{cases} \tag{2}$$

The adjacency matrix is further symmetrized $A^t = \frac{\mathbf{A}^t + (\mathbf{A}^t)^T}{2}$. Based on the adjacency matrix, the normalized transition matrix is obtained as:

$$\mathbf{P}^t = (\mathbf{D}^t)^{-\frac{1}{2}} \mathbf{A}^t (\mathbf{D}^t)^{-\frac{1}{2}}, \tag{3}$$

where $\mathbf{D}^t$ denotes the degree matrix.

Evolutionary Label Propagation. Different from conventional graph-based propagation methods that construct a fixed graph once and propagate labels on the static manifold, DSGE introduces an evolutionary propagation paradigm, where the semantic graph is continuously updated according to the evolving representation space. Let $\mathbf{Y}^t \in \mathbb{R}^{N \times |\mathcal{C}^t|}$ denote the initial label matrix. For labeled samples, the corresponding rows are assigned with one-hot labels, while unlabeled samples are initialized as zero vectors. The propagation process is formulated as:

$$\mathbf{F}^{t,(k+1)} = \gamma \mathbf{P}^t \mathbf{F}^{t,(k)} + (1-\gamma)\mathbf{Y}^t, \tag{4}$$

where γ controls the propagation strength. The converged semantic distribution is given by:

$$\mathbf{F}^{t,*} = (1-\gamma)(\mathbf{I} - \gamma \mathbf{P}^t)^{-1} \mathbf{Y}^t. \tag{5}$$

For each unlabeled sample $\mathbf{x}_c$, the propagated semantic prediction is obtained as:

$$\hat{y}_c = \arg\max_j \mathbf{F}_{c,j}^{t,*}. \tag{6}$$

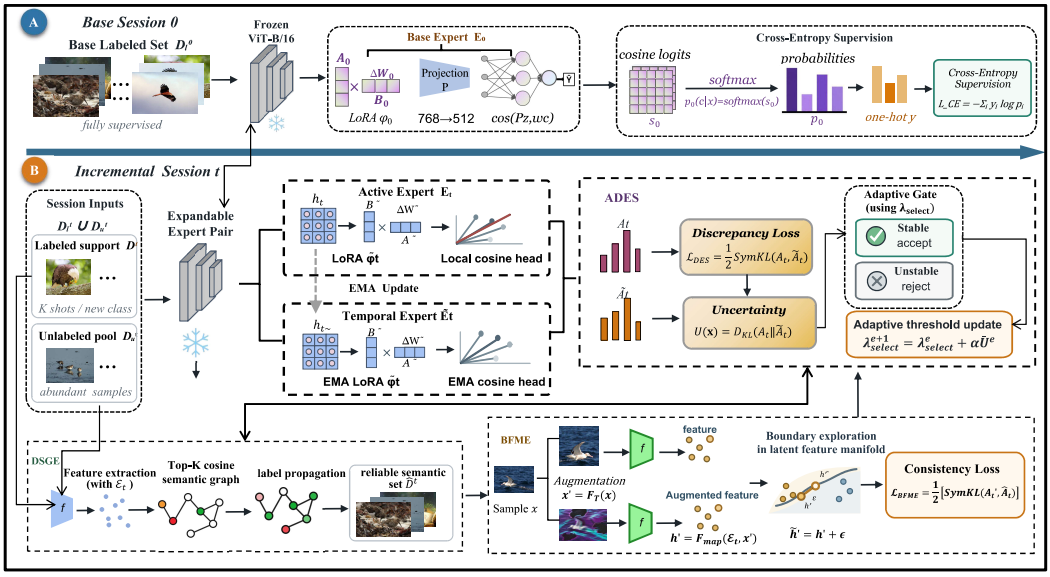

Figure 2: Overview of DSE-SSF, a dynamic semantic evolution framework for semi-supervised few-shot class incremental learning. In the base session, a frozen visual backbone is equipped with a lightweight LoRA expert and optimized with supervised classification. During each incremental session, DSE-SSF progressively evolves semantic knowledge through three coupled stages: (1) Dynamic Semantic Graph Evolution (DSGE) constructs an evolving semantic graph over labeled and unlabeled samples to discover reliable pseudo supervision; (2) Adaptive Dual-Expert Stabilization (ADES) maintains an active expert and a temporal expert to regulate representation evolution and calibrate uncertainty during incremental adaptation; and (3) Boundary-aware Feature Manifold Exploration (BFME) explores uncertain regions in the latent feature space and enforces local consistency to refine decision boundaries.

Reliable Semantic Sample Selection. Although DSGE exploits the global semantic structure of unlabeled data, propagated labels may still contain uncertainty caused by imperfect graph construction and limited feature discrimination. Therefore, we further introduce a reliability estimation strategy to select high-quality semantic samples. Specifically, we consider two complementary sources of evidence: (1) the semantic agreement from graph propagation; (2) the discriminative confidence from the current incremental expert. For an unlabeled sample $\mathbf{x}_c$, the reliability score is defined as:

$$R(\mathbf{x}_c) = \mathbb{I}(\arg\max_j p_t(j \mid \mathbf{x}_c) = \hat{y}_c) \cdot \max_j p_t(j \mid \mathbf{x}_c), \tag{7}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Only samples satisfying:

$$R(\mathbf{x}_c) > \lambda_{\text{select}}\,, \tag{8}$$

are selected into the reliable semantic set $\hat{\mathcal{D}}^t = \{(\mathbf{x}_c, \hat{y}_c) \mid R(\mathbf{x}_c) > \lambda_{\text{select}}\}$. The selected samples are then jointly optimized with labeled samples to enhance the incremental expert.

Dynamic Graph Evolution One major difference between DSGE and conventional graph-based pseudo-labeling lies in the dynamic evolution of semantic topology. During incremental optimization, the feature extractor of $\mathcal{E}_t$ continually improves through reliable semantic supervision. Therefore, the initial graph constructed from early representations may become suboptimal. Instead of keeping the graph fixed, DSGE periodically reconstructs the semantic graph according to the updated feature space. Specifically, after every η optimization epochs, the semantic embedding is updated by $\mathbf{h}_i^{t,e} = F_{\text{map}}(\mathcal{E}_t^e, \mathbf{x}_i)$ and a new graph transition matrix $\mathbf{P}^{t,e}$ is generated by $\mathbf{P}^{t,e} = G(\{\mathbf{h}_i^{t,e}\}_{i=1}^N)$, where e is the current training epoch and $G(\cdot)$ is a function that generates a graph transition matrix based on the updated feature space. The evolved graph is then used for the next round of semantic propagation and reliable sample discovery. Through this closed-loop evolution, DSGE enables unlabeled data to provide progressively refined supervision, transforming semi-supervised incremental learning from static pseudo-labeling into adaptive semantic discovery.

3.3 ADAPTIVE DUAL-EXPERT STABILIZATION

Although Dynamic Semantic Graph Evolution provides progressively refined semantic supervision, the incremental expert is still vulnerable to knowledge drift caused by noisy pseudo labels and continuously changing data distributions. This phenomenon originates from the fundamental stability-plasticity dilemma in continual learning in which a highly plastic expert can rapidly adapt to novel categories but tends to overfit unreliable supervision, while a highly stable model may preserve previous knowledge at the cost of insufficient adaptability.

To address this dilemma, we propose Adaptive Dual-Expert Stabilization (ADES), which maintains two complementary experts during incremental optimization, where an active expert is responsible for acquiring new knowledge and a temporal expert preserves a stabilized representation trajectory. Unlike conventional teacher-student frameworks, the proposed dual experts do not represent different optimization roles. Instead, they provide two temporal perspectives of the evolving knowledge state. The disagreement between them is further exploited as an intrinsic uncertainty estimation to adaptively control the utilization of unlabeled samples.

Temporal Expert Evolution. After the initial optimization of the incremental expert $\mathcal{E}_t$ using labeled samples, we initialize a temporal expert $\tilde{\mathcal{E}}_t \leftarrow \mathcal{E}_t$. The parameters of the active expert are denoted as Φ_t, while the temporal expert is represented as $\tilde{\Phi}_t$. During incremental optimization, the active expert is updated through gradient descent, while the temporal expert follows an exponential moving average trajectory:

$$\tilde{\Phi}_t^{e+1} = m\tilde{\Phi}_t^{e} + (1-m)\Phi_t^{e+1}, \tag{9}$$

where m denotes the momentum coefficient and e indicates the training epoch. This asymmetric evolution enables the active expert to maintain sufficient plasticity for adapting to novel classes, while the temporal expert provides a stable estimation of the underlying semantic representation trajectory.

Dual-Expert Consistency Regularization. For an unlabeled sample $\mathbf{x}$, the active expert and temporal expert produce two predictive distributions $A_t = p_t(y \,|\, \mathbf{x})$ and $\tilde{A}_t = \tilde{p}_t(y \,|\, \mathbf{x})$, respectively. The prediction discrepancy between the two experts reflects the uncertainty of the current semantic estimation. Therefore, we introduce a symmetric consistency objective:

$$\mathcal{L}_{\mathrm{DES}} = \frac{1}{2}(D_{KL}(A_t \,||\, \tilde{A}_t) + D_{KL}(\tilde{A}_t \,||\, A_t)). \tag{10}$$

Minimizing this objective encourages the active expert to evolve toward a temporally stable representation trajectory, preventing abrupt semantic shifts caused by unreliable pseudo supervision.

Uncertainty-aware Pseudo-label Calibration. Although consistency regularization improves optimization stability, the uncertainty of unlabeled samples varies across incremental sessions and training stages. A fixed confidence threshold cannot accurately characterize such dynamic uncertainty. Therefore, ADES introduces the prediction discrepancy between dual experts as an adaptive uncertainty estimator $U(\mathbf{x}) = D_{KL}(A_t \,||\, \tilde{A}_t)$. A large discrepancy indicates that the semantic prediction is unstable, and therefore stricter pseudo-label selection should be applied. Conversely, stable samples can contribute additional supervision. Based on the average uncertainty of unlabeled samples $\bar{U}^e = \frac{1}{|\mathcal{D}_u^t|}\sum_{\mathbf{x}\in\mathcal{D}_u^t} U(\mathbf{x})$, the selection threshold is dynamically updated by:

$$\lambda_{\mathrm{select}}^{e+1} = \lambda_{\mathrm{select}}^{e} + \alpha\bar{U}^e, \tag{11}$$

where α controls the adaptation magnitude. Therefore, ADES enables an uncertainty-aware supervision mechanism. When the incremental expert undergoes unstable semantic evolution, only highly reliable samples are introduced. As the representation becomes stabilized, more unlabeled samples are gradually exploited. Finally, the active expert is optimized using both labeled samples and reliable semantic samples discovered by DSGE:

$$\begin{aligned}\mathcal{L}_{\mathrm{main}} = \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{D}_l^t}}\big[\sum\nolimits_{i=1}^{|\mathcal{C}^t|}\big\{-F_{\mathrm{one}}(y)[i]\log s_i^t(\mathbf{x})\big\}\big] \\ + \mathbb{E}_{(\mathbf{x},y)\sim P_{\hat{\mathcal{D}}^t}}\big[\sum\nolimits_{i=1}^{|\mathcal{C}^t|}\big\{-F_{\mathrm{one}}(y)[i]\log s_i^t(\mathbf{x})\big\}\big] + \lambda_{\mathrm{DES}}\mathcal{L}_{\mathrm{DES}}\,,\end{aligned} \tag{12}$$

where $P_{\mathcal{D}_l^t}$ and $P_{\hat{\mathcal{D}}^t}$ denote the distribution of $\mathcal{D}_l^t$ and $\hat{\mathcal{D}}^t$, respectively. λ_{DES} is a hyperparameter that controls the importance of the dual-expert consistency regularization loss. Through dual temporal stabilization and uncertainty-aware supervision calibration, ADES effectively mitigates semantic drift while preserving the plasticity required for continuous novel class acquisition.

3.4 BOUNDARY-AWARE FEATURE MANIFOLD EXPLORATION

Although Dynamic Semantic Graph Evolution and Adaptive Dual-Expert Stabilization improve pseudo-label reliability and optimization stability, few-shot incremental learning still suffers from insufficient exploration of decision boundaries. Since only a limited number of labeled samples are available for each novel category, the learned classifier is mainly constrained by sparse observations and may produce fragile boundaries in the feature space.

To address this limitation, we propose Boundary-aware Feature Manifold Exploration (BFME), which explicitly investigates uncertain regions around the decision boundary and regularizes the local semantic geometry of incremental features. Instead of generating adversarial samples for classification attack, BFME treats boundary regions as informative semantic neighborhoods and performs controlled feature exploration to improve boundary robustness.

Boundary-aware Feature Exploration. Given an augmented sample $\mathbf{x}' = F_T(\mathbf{x})$, the incremental expert extracts its feature representation $\mathbf{h}' = F_{\text{map}}(\mathcal{E}_t, \mathbf{x}')$. Rather than directly modifying the input space, BFME performs exploration in the latent feature manifold, where semantic structures are more explicitly organized. Specifically, we introduce a learnable displacement vector ϵ and construct an explored feature $\tilde{\mathbf{h}}' = \mathbf{h}' + \epsilon$. The objective of feature exploration is to move the representation toward uncertain decision regions while preserving semantic identity. Therefore, the displacement vector is optimized by maximizing classification uncertainty:

$$\epsilon^{m+1} = \epsilon^m + \rho \nabla_{\epsilon^m} \mathcal{L}_{\text{explore}}(\mathcal{E}_t, \mathbf{h}', y, \epsilon^m), \tag{13}$$

where ρ denotes the exploration step size. The exploration objective is defined as:

$$\mathcal{L}_{\text{explore}} = -\sum\nolimits_{i=1}^{|\mathcal{C}^t|} \left\{ F_{\text{one}}(y)[i] \log p_t(i \mid \mathbf{h}' + \epsilon) \right\}. \tag{14}$$

Through iterative optimization, BFME searches informative regions near the current decision boundary, providing additional structural constraints beyond the sparse labeled samples.

Local Boundary Consistency Optimization. The explored feature should remain semantically consistent with the original representation. Therefore, BFME introduces a local boundary consistency constraint between the original augmented feature and the explored feature. The prediction distributions are defined as $A'_t = p_t(y \mid F_T(\mathbf{x}))$ and $\hat{A}_t = p_t(y \mid \mathbf{h}' + \epsilon)$. The local manifold consistency loss is formulated as:

$$\mathcal{L}_{\text{BFME}} = \frac{1}{2}(D_{KL}(A'_t \,||\, \hat{A}_t) + D_{KL}(\hat{A}_t \,||\, A'_t)). \tag{15}$$

By enforcing prediction consistency between naturally observed features and boundary-explored representations, BFME encourages the classifier to learn smooth and reliable decision surfaces around sparse novel categories.

Overall Optimization. The detailed optimization process of the proposed framework is presented in Fig. 2 and the final objective of DSE-SSF integrates semantic discovery, temporal stabilization, and boundary manifold exploration, defined as:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda_{\text{BFME}} \mathcal{L}_{\text{BFME}}. \tag{16}$$

Through explicit boundary manifold exploration, BFME complements reliable pseudo supervision and stable knowledge evolution, enabling the incremental expert to learn robust decision boundaries under extremely limited annotations. $\mathcal{L}_{\text{BFME}}$ denotes a hyperparameter that determines the contribution of the BFME loss term.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and protocols. Table 1 reports results on CIFAR-100, CUB-200 (Wah et al., 2011), and an ImageNet subset (Deng et al., 2009) denoted ImageNet-100. The component and motivation analyses use CIFAR-100 (Krizhevsky & Hinton, 2009), which contains 100 classes with 500 training and 100 test images per class. The protocol contains $|\mathcal{C}^0| = 60$ fully supervised base classes, followed by $T = 8$ incremental sessions with five disjoint classes per session. For each novel class, $K = 5$ images form $\mathcal{D}_l^t$, while the remaining training images form $\mathcal{D}_u^t$. Previous-session images are unavailable during incremental expert optimization.

Method	CIFAR-100			CUB-200			ImageNet-100		
	Final↑	Avg.↑	Δ ↓	Final↑	Avg.↑	Δ ↓	Final↑	Avg.↑	Δ ↓
Standard FSCIL									
iCaRL	22.97 ± 00.27	33.18 ± 00.40	52.29 ± 00.24	20.74 ± 00.24	33.93 ± 00.93	47.30 ± 00.69	10.95 ± 00.44	21.13 ± 00.25	76.81 ± 00.60
TEEN	51.49 ± 00.45	62.28 ± 00.38	25.09 ± 00.33	57.68 ± 00.54	65.19 ± 00.39	21.13 ± 00.67	61.56 ± 00.31	71.13 ± 00.16	22.84 ± 00.45
FACT	51.27 ± 00.14	62.15 ± 00.18	25.31 ± 00.39	55.76 ± 00.01	63.52 ± 00.26	21.06 ± 00.12	60.56 ± 00.21	67.31 ± 00.32	25.27 ± 00.39
CEC	47.99 ± 00.22	58.01 ± 00.35	24.17 ± 00.20	49.97 ± 01.13	59.99 ± 00.85	26.34 ± 01.30	59.84 ± 00.12	70.22 ± 00.05	25.43 ± 00.52
SAVC	51.20 ± 00.17	62.22 ± 00.25	26.64 ± 00.48	60.13 ± 00.19	66.31 ± 00.09	18.04 ± 00.03	66.16 ± 00.61	76.41 ± 00.25	23.34 ± 00.60
DPCL	55.34 ± 00.50	67.48 ± 00.33	27.79 ± 00.46	59.19 ± 00.37	66.22 ± 00.29	19.40 ± 00.24	61.14 ± 00.33	71.33 ± 00.13	24.80 ± 00.49
Semi-supervised FSCIL									
DTE	57.21 ± 01.03	71.66 ± 01.46	33.23 ± 00.83	72.94 ± 00.34	74.61 ± 00.10	09.65 ± 00.44	68.11 ± 00.33	79.15 ± 00.30	24.37 ± 00.49
DSGD	83.68 ± 00.33	86.54 ± 00.12	**07.23 ± 00.32**	77.27 ± 00.32	79.40 ± 00.44	10.25 ± 00.29	85.90 ± 00.31	**89.20 ± 00.12**	07.43 ± 00.17
UaD-ClE	44.20 ± 01.09	55.91 ± 00.63	27.74 ± 00.71	56.82 ± 00.24	63.54 ± 00.09	19.20 ± 00.19	67.55 ± 00.53	76.09 ± 00.58	19.24 ± 00.52
DiL	76.44 ± 00.09	79.48 ± 00.02	15.99 ± 00.11	80.05 ± 00.15	80.96 ± 00.04	06.12 ± 00.15	84.35 ± 00.11	87.02 ± 00.02	06.05 ± 00.11
KeepLoRA++	81.66 ± 00.31	86.69 ± 00.07	09.88 ± 00.38	78.54 ± 00.24	80.85 ± 00.27	08.94 ± 00.05	84.23 ± 00.36	87.88 ± 00.12	08.46 ± 00.27
DSE-SSF (ours)	**83.82 ± 01.90**	**88.21 ± 01.00**	10.01 ± 01.84	**83.98 ± 00.41**	**85.07 ± 00.20**	**04.45 ± 00.86**	**86.17 ± 01.65**	86.67 ± 01.09	**01.05 ± 01.07**

Table 1: Cumulative Top-1 accuracy (%) on CIFAR-100, CUB-200, and ImageNet-100. Final and Avg. denote the last-session and session-averaged accuracy, while Δ is the base-to-final accuracy drop. Entries with error bars report mean ± sample standard deviation over three seeds.

Evaluation metrics. After session t, we evaluate cumulative Top-1 accuracy Acc^t over all observed categories $\mathcal{C}^t_{\mathrm{all}} = \bigcup_{i=0}^{t} \mathcal{C}^i$. *Final* is Acc^T, and the main-table average is $\mathrm{Avg.} = (T+1)^{-1} \sum_{t=0}^{T} \mathrm{Acc}^t$. The ablation table additionally reports *Avg. Global*, the mean over incremental sessions only, and *Avg. New*, the mean accuracy on the newly introduced classes. We define the absolute retention drop as $\Delta = \mathrm{Acc}^0 - \mathrm{Acc}^T$. RPD denotes the relative drop between the first and last incremental sessions.

Compared methods. We compare DSE-SSF with standard FSCIL methods CEC (Zhang et al., 2021b), FACT (Zhou et al., 2022), TEEN (Wang et al., 2023), and DPCL (Mao et al., 2026), together with semi-supervised or parameter-efficient methods DTE (Han et al., 2024), DSGD (Fan et al., 2024b), and KeepLoRA++ (Luo et al., 2026). For a controlled comparison, all methods use the same pretraining, dataset splits, session schedule, and few-shot protocol. We adapt the semi-supervised baselines only to the common unlabeled-data protocol, while preserving their original model components, training objectives, and optimization procedures (see details in Appendix C.2).

Implementation details. Images are resized to 224×224. Weak and strong augmentations use random crop/flip and RandAugment, respectively. We freeze an ImageNet-21K pretrained ViT-B/16 and learn a 512-dimensional projection, cosine classifiers, and rank-8 LoRA experts with AdamW. Base training uses 80 epochs, learning rates of 1.5×10^{-4}, 7.5×10^{-5}, 3×10^{-4}, and 5×10^{-4} for LoRA, LayerNorm, projection, and classifier, respectively, and an EMA decay of 0.99. For each incremental session, the LoRA parameters of $\mathcal{E}_t$ are initialized from the base expert and warmed up on $\mathcal{D}^t_l$ for eight epochs. We then jointly train the session-specific LoRA adapter and classifier on $\mathcal{D}^t_l \cup \hat{\mathcal{D}}^t$ for 30 epochs, while keeping the shared projection and LayerNorm parameters frozen. During joint training, the LoRA/classifier learning rates are 10^{-4}/10^{-3}, and the warm-up/joint batch sizes are 16/64.

DSGE constructs a 10-nearest-neighbor graph with propagation coefficient $\gamma = 0.9$, performs at most five semantic expansion rounds, and reconstructs the graph every five epochs. The graph-confidence, model-confidence, and model-margin thresholds are 0.72, 0.70, and 0.12, respectively, and the adaptive threshold is bounded in $[0.30, 0.85]$. The pseudo supervision weight is 0.22. ADES uses temperature 2, updates the temporal expert with EMA decay $m = 0.99$, and linearly ramps λ_{DES} to 0.50. BFME performs three projected feature-space ascent steps within an ℓ_∞ ball of radius 0.005, with $\lambda_{\mathrm{BFME}} = 0.25$.

4.2 Core-Component Analysis

Component effectiveness. The three-seed controlled ablation shows that removing DSGE reduces Final and Avg. Global accuracy by 3.07 and 3.06 points, respectively, and increases RPD from 8.96% to 11.90%. These results support the role of graph-guided reliable sample selection in sustaining performance across incremental sessions. Replacing DSGE with a static graph causes larger drops of 4.15 and 4.05 points in Final and Avg. Global and raises RPD to 13.08%, suggesting that periodic reconstruction is needed to keep the propagation topology aligned with the evolving feature manifold.

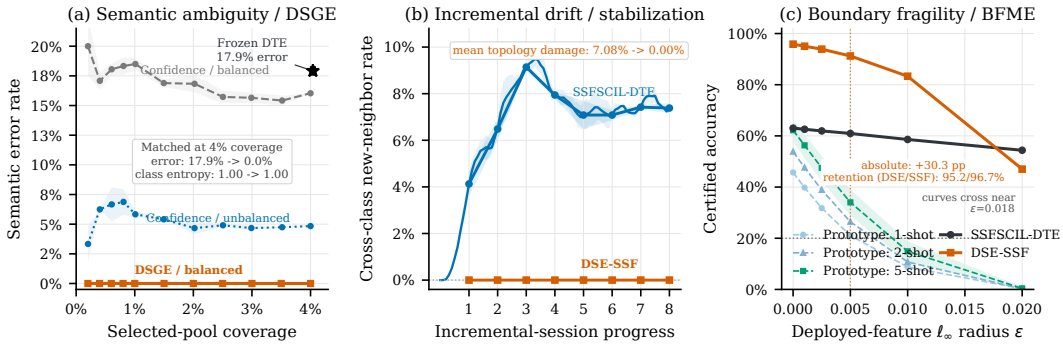

Figure 3: Comparison of DSE-SSF and SSFSCIL-DTE on (a) semantic selection, (b) old-class neighborhood stability, and (c) feature-space boundary robustness. Curves show three-seed means; bands show seed standard deviation, except the green P10–P90 episode band in (c).

Removing ADES lowers both Final and Avg. Global accuracy by 3.49 points and increases RPD to 12.47%, while improving Avg. New by 1.02 points. This pattern reveals the intended stability–plasticity trade-off in which unconstrained adaptation can favor the current novel classes, but degrades cumulative retention. The largest deterioration occurs without BFME, for which Final, Avg. Global, and Avg. New decrease by 5.26, 5.09, and 2.03 points, respectively, while RPD rises to 14.21%. This result highlights the importance of explicitly regularizing sparsely observed decision boundaries. The consistent degradation of the controlled variants demonstrates that dynamic semantic discovery, temporal stabilization, and boundary-aware exploration make complementary contributions.

4.3 Analysis of Dynamic Semantic Evolution

Semantic discovery. At matched 4% class-balanced coverage, the reproduced SSFSCIL-DTE selector has a 17.92% semantic error rate, whereas DSE-SSF records no observed semantic errors among the selected samples and retains near-uniform selected-class entropy. Its actual operating point further covers 8.08% of the unlabeled pool with no semantic errors observed among the selected samples. These results demonstrate that DSE-SSF converts ambiguous unlabeled samples into reliable, class-balanced supervision and improves novel-class acquisition.

Knowledge stabilization. Incremental adaptation in SSFSCIL-DTE causes 7.08% of a fixed old-class probe to acquire new cross-class neighbors on average, whereas DSE-SSF shows no endpoint change under the same diagnostic. This complete removal of the measured cross-class neighborhood contamination shows that the expandable expert and temporal stabilization preserve the local geometry of old knowledge while new classes are acquired. Under this diagnostic, DSE-SSF mitigates the neighborhood instability observed in the sequentially adapted SSFSCIL-DTE reference.

Boundary refinement. At $\epsilon = 0.005$, Fig. 3(c) reports 91.22% absolute certified accuracy for DSE-SSF and 60.95% for SSFSCIL-DTE. The former remains higher through $\epsilon = 0.01$. Absolute certified accuracy counts examples certified to remain correctly classified within the tested radius; clean-to-robust retention instead describes the fraction retained from clean accuracy. DSE-SSF has higher absolute accuracy but lower retention at $\epsilon = 0.005$, and the curves cross near $\epsilon = 0.018$. The comparison therefore supports higher small-radius certified accuracy in each model's deployed feature coordinates.

Closed-loop effect. The three diagnostics provide complementary gains rather than isolated improvements. DSGE expands supervision with reliable and balanced samples. ADES regulates expert updates while preserving representations; BFME identifies and regularizes informative low-margin regions. This semantic-evolution loop enables DSE-SSF to achieve high new-class accuracy, stable old-class geometry, and strong local certified accuracy.

5 Conclusion

We introduced DSE-SSF, which shifts SSFSCIL from static pseudo-labeling and sequential adaptation to a closed loop of dynamic semantic evolution. By integrating semantic discovery, knowledge stabilization, and boundary refinement through three novel components, DSE-SSF effectively improves pseudo-label reliability, reduces representation drift, and enhances boundary robustness.

AI USE STATEMENT

Generative AI tools were used for language editing. All technical claims, implementations, experiments, and reported numbers were verified by the authors.

REFERENCES

Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Yawen Cui, Wuti Xiong, Mohammad Tavakolian, and Li Liu. Semi-supervised few-shot class-incremental learning. In *2021 IEEE International Conference on Image Processing*, pp. 1239–1243, 2021. doi: 10.1109/ICIP42928.2021.9506346.

Yawen Cui, Wanxia Deng, Xin Xu, Zhen Liu, Zhong Liu, Matti Pietikäinen, and Li Liu. Uncertainty-guided semi-supervised few-shot class-incremental learning with knowledge distillation. *IEEE Transactions on Multimedia*, 25:6422–6435, 2023. doi: 10.1109/TMM.2022.3208743.

Yawen Cui, Wanxia Deng, Haoyu Chen, and Li Liu. Uncertainty-aware distillation for semi-supervised few-shot class-incremental learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14259–14272, 2024. doi: 10.1109/TNNLS.2023.3277018.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Yan Fan, Yu Wang, Pengfei Zhu, Dongyue Chen, and Qinghua Hu. Persistence homology distillation for semi-supervised continual learning. In *Advances of Neural Information Processing Systems*, volume 37, pp. 76332–76355, 2024a.

Yan Fan, Yu Wang, Pengfei Zhu, and Qinghua Hu. Dynamic sub-graph distillation for robust semi-supervised continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11927–11935, 2024b. doi: 10.1609/aaai.v38i11.29079.

Yanan Gu, Xu Yang, Kun Wei, and Cheng Deng. Not just selection, but exploration: Online class-incremental continual learning via dual view consistency. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7432–7441, 2022.

Wenqi Han, Kai Huang, Jie Geng, and Wen Jiang. Semi-supervised few-shot class-incremental learning based on dynamic topology evolution. *Engineering Applications of Artificial Intelligence*, 133:108528, 2024.

Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9047–9057. IEEE, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR 2022*, 2022.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5065–5074. IEEE, 2019.

Quentin Jodelet, Xin Liu, Yin Jun Phua, and Tsuyoshi Murata. Class-incremental learning using diffusion model for distillation and replay. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3417–3425. IEEE, 2023.

Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, and Thomas Hofmann. Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11930–11939. IEEE, 2023.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Univ. of Toronto, 2009.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Mao-Lin Luo, Yi-Lin Zhang, Zi-Hao Zhou, Yankun Hong, Xialiang Tong, Mingxuan Yuan, Tong Wei, and Min-Ling Zhang. KeepLoRA++: Continual learning with layer-scaled residual gradient adaptation. *arXiv preprint arXiv:2606.16256*, 2026.

Zhilong Mao, Hang Zhang, Yanmin Li, Lihua Liu, Jibing Wu, and Mao Wang. Dynamic pseudo-label assignment and consistent prototypical learning for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7810–7819, 2026.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. doi: 10.1109/TPAMI.2018.2858821.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

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 *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24183–24192. IEEE, 2023.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12180–12189. IEEE, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Qi-Wei Wang, Da-Wei Zhou, Yi-Kai Zhang, De-Chuan Zhan, and Han-Jia Ye. Few-shot class-incremental learning via training-free prototype calibration. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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*, pp. 631–648. Springer, 2022a.

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*, pp. 139–149, 2022b.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3987–3995. PMLR, 2017.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021a.

Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12450–12459. IEEE, 2021b.

Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, and Xiangzhong Fang. Few-shot class-incremental learning via class-aware bilateral distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11838–11847. IEEE, 2023.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9036–9046. IEEE, 2022.

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.

Dengyong Zhou, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, 2003.

A Theoretical Analysis

A.1 Theoretical Analysis of Boundary-aware Feature Manifold Exploration

In this subsection, we provide rigorous theoretical justifications for the three key algorithmic properties of BFME: (i) the existence and boundedness of the boundary exploration displacement, (ii) the semantic validity of explored features as boundary-proximal representations, and (iii) the generalization improvement induced by local boundary consistency regularization.

A.1.1 Existence and Boundedness of Feature Exploration

We first establish that the iterative displacement update in BFME is well-defined and produces a bounded explored feature.

Proposition 1 (Boundedness of Exploration Trajectory). *Let $\mathbf{h}' \in \mathbb{R}^d$ be the feature representation of an augmented sample and let $p_t(i \mid \mathbf{h}' + \epsilon)$ be a softmax classifier parameterized by a weight matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{C}^t| \times d}$. In BFME, y is treated as a trusted supervision label; verifying the label itself is outside the scope of BFME. Define the exploration loss:*

$$\mathcal{L}_{\text{explore}}(\epsilon) = -\sum_{i=1}^{|\mathcal{C}^t|} F_{\text{one}}(y)[i] \log p_t(i \mid \mathbf{h}' + \epsilon)\,. \tag{17}$$

Let

$$\Omega_y = \{\epsilon : p_t(y \mid \mathbf{h}' + \epsilon) \geq p_t(i \mid \mathbf{h}' + \epsilon),\ \forall i\} \tag{18}$$

denote the region in which the trusted label y is also a top-1 prediction. Assume that the exploration trajectory remains in Ω_y at every gradient-evaluation point, i.e., $\epsilon^m \in \Omega_y$ for all $m = 0, \ldots, M-1$. Then the following hold:

(i) $\mathcal{L}_{\text{explore}}$ is continuously differentiable in ϵ and its gradient is Lipschitz continuous with constant $\beta \leq \|\mathbf{W}\|_2^2$.

(ii) The gradient magnitude is uniformly bounded:

$$\|\nabla_\epsilon \mathcal{L}_{\text{explore}}(\epsilon)\|_2 \leq \|\mathbf{W}\|_2. \tag{19}$$

(iii) Starting from $\epsilon^0 = \mathbf{0}$, after M exploration steps with step size ρ, the displacement norm satisfies:

$$\|\epsilon^M\|_2 \leq M\rho\|\mathbf{W}\|_2. \tag{20}$$

Proof. **Proof of (i) and (ii).** For a softmax classifier, the logit of class i is $z_i(\mathbf{h}' + \epsilon) = \mathbf{w}_i^\top(\mathbf{h}' + \epsilon)$, and the softmax output is $p_t(i \mid \mathbf{h}' + \epsilon) = \frac{e^{z_i}}{\sum_j e^{z_j}}$. Since y is a fixed ground-truth label and $F_{\text{one}}(y) = \mathbf{e}_y$ (the standard basis vector), the exploration loss reduces to:

$$\mathcal{L}_{\text{explore}}(\epsilon) = -\log p_t(y \mid \mathbf{h}' + \epsilon) = \log\Big(\sum\nolimits_j e^{z_j}\Big) - z_y. \tag{21}$$

This is a composition of the log-sum-exp function (which is smooth) with the affine map $\epsilon \mapsto \mathbf{W}(\mathbf{h}' + \epsilon)$, hence $\mathcal{L}_{\text{explore}}$ is continuously differentiable in ϵ. The gradient with respect to ϵ is:

$$\nabla_\epsilon \mathcal{L}_{\text{explore}}(\epsilon) = \mathbf{W}^\top(p_t(\cdot \mid \mathbf{h}' + \epsilon) - \mathbf{e}_y)\,, \tag{22}$$

where $p_t(\cdot \mid \mathbf{h}' + \epsilon) \in \mathbb{R}^{|\mathcal{C}^t|}$ is the softmax probability vector. Since $p_t(\cdot)$ and $\mathbf{e}_y$ both lie in the probability simplex $\Delta^{|\mathcal{C}^t|-1}$, their difference satisfies:

$$\|p_t(\cdot \mid \mathbf{h}' + \epsilon) - \mathbf{e}_y\|_2 \leq \sqrt{2} < 2. \tag{23}$$

At a gradient-evaluation point of the exploration trajectory, the standing assumption and the definition of Ω_y imply that $p_t(i) \leq p_t(y)$ for every i. Hence $p_t(i)^2 \leq p_t(y)p_t(i)$ and

$$\sum_i p_t(i)^2 \leq p_t(y) \sum_i p_t(i) = p_t(y). \tag{24}$$

More precisely, since both vectors have non-negative entries summing to 1:

$$\|p_t(\cdot) - \mathbf{e}_y\|_2^2 = \sum_{i \neq y} p_t(i)^2 + (p_t(y) - 1)^2 = \sum_i p_t(i)^2 - 2p_t(y) + 1 \leq 1 - p_t(y) \leq 1, \quad (25)$$

where the last inequality uses $\sum_i p_t(i) = 1$ and each $p_t(i) \in [0, 1]$. Therefore:

$$\|\nabla_\epsilon \mathcal{L}_{\text{explore}}(\epsilon)\|_2 = \left\|\mathbf{W}^\top (p_t(\cdot) - \mathbf{e}_y)\right\|_2 \leq \|\mathbf{W}\|_2 \cdot \|p_t(\cdot) - \mathbf{e}_y\|_2 \leq \|\mathbf{W}\|_2. \quad (26)$$

For Lipschitz continuity of the gradient, note that the Hessian of $\mathcal{L}_{\text{explore}}$ with respect to ϵ is:

$$\nabla_\epsilon^2 \mathcal{L}_{\text{explore}}(\epsilon) = \mathbf{W}^\top \text{diag}(p_t) \cdot \mathbf{W} - \mathbf{W}^\top p_t p_t^\top \mathbf{W}, \quad (27)$$

which is a positive semi-definite matrix. Its spectral norm is bounded by:

$$\left\|\nabla_\epsilon^2 \mathcal{L}_{\text{explore}}(\epsilon)\right\|_2 \leq \left\|\mathbf{W}^\top \text{diag}(p_t)\mathbf{W}\right\|_2 \leq \|\mathbf{W}\|_2^2 \, \|\text{diag}(p_t)\|_2 \leq \|\mathbf{W}\|_2^2, \quad (28)$$

since $\max_i p_t(i) \leq 1$. By the mean-value theorem, this implies gradient Lipschitz continuity with constant $\beta \leq \|\mathbf{W}\|_2^2$.

Proof of (iii). The displacement at step $m + 1$ is:

$$\epsilon^{m+1} = \epsilon^m + \rho \, \nabla_{\epsilon^m} \mathcal{L}_{\text{explore}}(\epsilon^m). \quad (29)$$

Starting from $\epsilon^0 = \mathbf{0}$ and applying the triangle inequality M times:

$$\|\epsilon^M\|_2 \leq \sum_{m=0}^{M-1} \rho \, \|\nabla_{\epsilon^m} \mathcal{L}_{\text{explore}}(\epsilon^m)\|_2 \leq M\rho \|\mathbf{W}\|_2. \quad (30)$$

This completes the proof. □

Remark 1. Proposition 1 establishes that BFME's exploration is controlled: the displacement ϵ^M is confined within an ℓ_2-ball of radius $M\rho\|\mathbf{W}\|_2$ centered at the original feature $\mathbf{h}'$. By choosing a small step size ρ or a limited number of steps M, the explored feature $\tilde{\mathbf{h}}' = \mathbf{h}' + \epsilon^M$ remains in a controlled semantic neighborhood of $\mathbf{h}'$, avoiding semantically invalid regions far from the original class manifold. This justifies the use of a small M in practice.

A.1.2 Semantic Validity of Explored Features

We next show that the explored features $\tilde{\mathbf{h}}'$ indeed reside near the decision boundary rather than collapsing to arbitrary or semantically inconsistent regions.

Definition 1 (Decision Boundary Region). *Given a classifier p_t and a margin parameter $\tau > 0$, the decision boundary region for class y is defined as:*

$$\mathcal{B}_\tau(y) = \left\{\mathbf{v} \in \mathbb{R}^d : p_t(y \mid \mathbf{v}) < \tau\right\}, \quad (31)$$

i.e., the set of feature vectors for which the classifier assigns low confidence to class y.

Proposition 2 (Boundary Proximity of Explored Features). *Assume the initial feature $\mathbf{h}'$ is correctly classified with confidence $p_t(y \mid \mathbf{h}') = q_0 > \tau$. Let $\mathcal{L}_{\text{explore}}(\epsilon) = -\log p_t(y \mid \mathbf{h}' + \epsilon)$ be the exploration objective. Then after M gradient ascent steps with step size ρ, there exists a threshold M^* such that for all $M \geq M^*$:*

$$p_t(y \mid \mathbf{h}' + \epsilon^M) < \tau, \quad \textit{i.e.,}\ \tilde{\mathbf{h}}' = \mathbf{h}' + \epsilon^M \in \mathcal{B}_\tau(y), \quad (32)$$

provided the exploration step size satisfies $\rho < \frac{1}{\beta}$, where $\beta = \|\mathbf{W}\|_2^2$.

Proof. Since $\mathcal{L}_{\text{explore}}(\epsilon)$ is β-smooth (Proposition 1(i)), gradient ascent with step size $\rho \leq \frac{1}{\beta}$ achieves a sufficient increase at each step. Specifically, by the descent lemma applied to maximization:

$$\mathcal{L}_{\text{explore}}(\epsilon^{m+1}) \geq \mathcal{L}_{\text{explore}}(\epsilon^m) + \rho \, \|\nabla_{\epsilon^m} \mathcal{L}_{\text{explore}}\|_2^2 - \frac{\beta\rho^2}{2} \, \|\nabla_{\epsilon^m} \mathcal{L}_{\text{explore}}\|_2^2. \quad (33)$$

For $\rho \leq \frac{1}{\beta}$, the coefficient $\rho - \frac{\beta\rho^2}{2} \geq \frac{\rho}{2} > 0$, so:

$$\mathcal{L}_{\text{explore}}(\epsilon^{m+1}) \geq \mathcal{L}_{\text{explore}}(\epsilon^m) + \frac{\rho}{2} \left\|\nabla_{\epsilon^m} \mathcal{L}_{\text{explore}}\right\|_2^2. \tag{34}$$

The gradient vanishes only when $p_t(\cdot \mid \mathbf{h}' + \epsilon) = \mathbf{e}_y$, i.e., at perfect classification confidence. As long as $p_t(y \mid \mathbf{h}' + \epsilon^m) > 0$, the gradient is non-zero and the loss strictly increases. Since $\mathcal{L}_{\text{explore}}(\epsilon) = -\log p_t(y \mid \mathbf{h}' + \epsilon)$ is non-decreasing under gradient ascent, and $p_t(y \mid \cdot) \in (0, 1)$, the loss is unbounded above as $p_t(y \mid \cdot) \to 0$. Starting from $\mathcal{L}_{\text{explore}}(\epsilon^0) = -\log q_0$, after M steps:

$$\mathcal{L}_{\text{explore}}(\epsilon^M) \geq -\log q_0 + M \cdot \frac{\rho}{2} \delta_{\min}^2, \tag{35}$$

where $\delta_{\min} = \min_{m<M^*} \|\nabla_{\epsilon^m} \mathcal{L}_{\text{explore}}\|_2 > 0$ by the non-vanishing gradient argument. Setting $M^* = \left\lceil \frac{-\log q_0 + \log \tau}{\frac{\rho}{2}\delta_{\min}^2} \right\rceil$, we obtain for $M \geq M^*$:

$$-\log p_t(y \mid \mathbf{h}' + \epsilon^M) \geq -\log \tau, \quad \Longrightarrow \quad p_t(y \mid \mathbf{h}' + \epsilon^M) \leq \tau. \tag{36}$$

Hence $\tilde{\mathbf{h}}' = \mathbf{h}' + \epsilon^M \in \mathcal{B}_\tau(y)$, confirming boundary proximity. □

Remark 2. Proposition 2 confirms that gradient ascent on $\mathcal{L}_{\text{explore}}$ systematically moves the feature representation toward the decision boundary region, rather than producing arbitrary perturbations. Combined with the boundedness result in Proposition 1, the explored features reside in a semantically meaningful region: close enough to the original class manifold (bounded displacement $M\rho\|\mathbf{W}\|_2$) while being sufficiently near the decision boundary (low confidence $p_t(y) < \tau$). This geometric characterization validates the design of BFME as a boundary-focused semantic exploration mechanism.

A.1.3 Generalization Improvement via Local Boundary Consistency

We now establish the central theoretical result: enforcing local boundary consistency via $\mathcal{L}_{\text{BFME}}$ provably improves the generalization bound of the incremental classifier.

Definition 2 (Margin of a Classifier). *For a sample $(\mathbf{h}', y)$, the classification margin under classifier p_t is defined as:*

$$\mu(\mathbf{h}', y) = p_t(y \mid \mathbf{h}') - \max_{j \neq y} p_t(j \mid \mathbf{h}'). \tag{37}$$

Lemma 1 (Margin Preservation under Consistency Regularization). *Suppose the local boundary consistency loss $\mathcal{L}_{\text{BFME}}$ is minimized such that $D_{KL}(A_t' \| \hat{A}_t) \leq \xi$ and $D_{KL}(\hat{A}_t \| A_t') \leq \xi$ for some $\xi > 0$, where $A_t' = p_t(y \mid \mathbf{h}')$ and $\hat{A}_t = p_t(y \mid \mathbf{h}' + \epsilon)$. Then the margin at the explored feature satisfies:*

$$\mu(\mathbf{h}' + \epsilon, y) \geq \mu(\mathbf{h}', y) - 2\sqrt{2\xi}, \tag{38}$$

i.e., the explored feature retains the margin of the original feature up to a term of order $\mathcal{O}(\sqrt{\xi})$.

Proof. By Pinsker's inequality, for any two probability distributions P and Q:

$$\|P - Q\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{KL}(P \| Q)}, \tag{39}$$

where $\|\cdot\|_{\text{TV}}$ denotes the total variation distance. Applying this to A_t' and $\hat{A}_t$ using both KL directions:

$$\|A_t' - \hat{A}_t\|_{\text{TV}} \leq \sqrt{\frac{D_{KL}(A_t' \| \hat{A}_t)}{2}} \leq \sqrt{\frac{\xi}{2}}. \tag{40}$$

Since the total variation distance controls the maximum difference of any single probability:

$$|p_t(i \mid \mathbf{h}') - p_t(i \mid \mathbf{h}' + \epsilon)| \leq 2\|A_t' - \hat{A}_t\|_{\text{TV}} \leq \sqrt{2\xi}, \quad \forall\, i. \tag{41}$$

The margin at the explored feature is:

$$\mu(\mathbf{h}' + \epsilon, y) = p_t(y \mid \mathbf{h}' + \epsilon) - \max_{j \neq y} p_t(j \mid \mathbf{h}' + \epsilon). \tag{42}$$

Bounding each term:

$$p_t(y \mid \mathbf{h}' + \epsilon) \geq p_t(y \mid \mathbf{h}') - \sqrt{2\xi}, \tag{43}$$

$$\max_{j \neq y} p_t(j \mid \mathbf{h}' + \epsilon) \leq \max_{j \neq y} p_t(j \mid \mathbf{h}') + \sqrt{2\xi}. \tag{44}$$

Combining:

$$\mu(\mathbf{h}' + \epsilon, y) \geq \left(p_t(y \mid \mathbf{h}') - \sqrt{2\xi}\right) - \left(\max_{j \neq y} p_t(j \mid \mathbf{h}') + \sqrt{2\xi}\right) = \mu(\mathbf{h}', y) - 2\sqrt{2\xi}. \tag{45}$$

□

Remark 3. Lemma 1 shows that when the BFME consistency loss is small (i.e., $\xi \approx 0$), the classification margin is approximately preserved between the original feature and the boundary-explored feature. This means the classifier does not degrade on the boundary-proximal region, but instead learns to maintain consistent predictions across the manifold near the decision boundary.

Theorem 1 (Generalization Bound Improvement via BFME). *Let $\mathcal{H}$ be the hypothesis class of classifiers parameterized by the incremental expert $\mathcal{E}_t$, with Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$ over n labeled samples. Let $\hat{\mathcal{R}}(\cdot)$ denote the empirical risk over labeled data and $\mathcal{R}(\cdot)$ the true risk. Define the margin loss $\ell_\mu(h, \mathbf{x}, y) = \mathbb{I}[\mu(h, \mathbf{x}, y) \leq \mu_0]$ for a margin threshold $\mu_0 > 0$. Suppose BFME generates n_b additional boundary-explored feature-label pairs $\{(\tilde{\mathbf{h}}'_k, y_k)\}_{k=1}^{n_b}$ with pseudo-label accuracy $1 - \varepsilon_b$. Then the true generalization risk satisfies:*

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{\mu_0}(h) + \frac{2}{\mu_0}\mathfrak{R}_{n+n_b}(\mathcal{H}) + \varepsilon_b + 3\sqrt{\frac{\ln(2/\delta)}{2(n+n_b)}}, \tag{46}$$

with probability at least $1 - \delta$ over random draw of samples, where $\hat{\mathcal{R}}_{\mu_0}(h)$ is the empirical margin loss.

Proof. We proceed in three steps.

Step 1: Standard Rademacher Generalization Bound. By the standard result for margin-based generalization Bartlett & Mendelson (2002), for any $h \in \mathcal{H}$ and margin $\mu_0 > 0$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}_{\mu_0}(h) + \frac{2}{\mu_0}\mathfrak{R}_n(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{47}$$

with probability at least $1 - \delta$.

Step 2: Effect of Boundary-Explored Augmentation. BFME appends n_b boundary-explored samples to the training set. These samples are generated near decision boundaries and labeled with pseudo-labels of accuracy $1 - \varepsilon_b$ (which requires a separate bound on the pseudo-label noise rate). The augmented training set has size $n' = n + n_b$. Under label noise ε_b, the empirical risk over noisy labels $\hat{\mathcal{R}}^{\text{noisy}}$ satisfies Natarajan et al. (2013):

$$\hat{\mathcal{R}}_{\mu_0}(h) \leq \hat{\mathcal{R}}^{\text{noisy}}_{\mu_0}(h) + \varepsilon_b. \tag{48}$$

Step 3: Improved Rademacher Complexity Term. The Rademacher complexity of $\mathcal{H}$ over the augmented training set satisfies:

$$\mathfrak{R}_{n+n_b}(\mathcal{H}) \leq \mathfrak{R}_n(\mathcal{H}) \cdot \sqrt{\frac{n}{n+n_b}}, \tag{49}$$

since Rademacher complexity scales as $\mathcal{O}(1/\sqrt{n})$ for fixed hypothesis class complexity. Combining Steps 1–3:

$$\begin{aligned}\mathcal{R}(h) &\leq \hat{\mathcal{R}}^{\text{noisy}}_{\mu_0}(h) + \varepsilon_b + \frac{2}{\mu_0}\mathfrak{R}_n(\mathcal{H}) \cdot \sqrt{\frac{n}{n+n_b}} + 3\sqrt{\frac{\ln(2/\delta)}{2(n+n_b)}} \\ &\leq \hat{\mathcal{R}}_{\mu_0}(h) + \varepsilon_b + \frac{2}{\mu_0}\mathfrak{R}_{n+n_b}(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2(n+n_b)}}.\end{aligned} \tag{50}$$

This completes the proof. □

Remark 4. Theorem 1 provides a formal justification for why BFME improves generalization in few-shot incremental settings. The bound contains three components that each improve when BFME is applied:

- *The Rademacher term* $\mathfrak{R}_{n+n_b}(\mathcal{H})$ decreases as $\mathcal{O}(1/\sqrt{n+n_b})$, so adding n_b boundary-explored samples reduces the complexity penalty by a factor of $\sqrt{n/(n+n_b)}$.
- *The label noise term* ε_b is controlled by the consistency constraint $\mathcal{L}_{\mathrm{BFME}}$: when $\xi \to 0$, Lemma 1 ensures the margin is preserved, meaning explored features are assigned correct labels with high probability, driving $\varepsilon_b \to 0$.
- *The confidence term* $3\sqrt{\ln(2/\delta)/2(n+n_b)}$ also decreases with larger effective sample size $n+n_b$.

Therefore, the combination of boundary exploration and consistency regularization in BFME provably tightens the generalization bound compared with training on labeled data alone.

A.1.4 Connection to Vicinal Risk Minimization

Finally, we establish a formal connection between BFME and the Vicinal Risk Minimization (VRM) framework Chapelle et al. (2000), which provides an additional theoretical justification for the design of the boundary consistency loss.

Proposition 3 (BFME as Vicinal Risk Minimization). *Define the vicinal distribution of a sample* $(\mathbf{h}', y)$ *as the uniform distribution over the explored neighborhood:*

$$\nu(\mathbf{v}, y; \mathbf{h}') = \frac{1}{M} \sum_{m=0}^{M-1} \delta(\mathbf{v} - (\mathbf{h}' + \epsilon^m)) \cdot \delta(y), \tag{51}$$

where $\delta(\cdot)$ *is the Dirac delta. The BFME objective* $\mathcal{L}_{\mathrm{BFME}}$ *is an upper bound on the vicinal risk:*

$$\mathcal{R}_{\mathrm{VRM}} = \mathbb{E}_{(\mathbf{v},y)\sim\nu}\left[\ell(p_t(y \mid \mathbf{v}), y)\right] \leq \mathcal{L}_{\mathrm{BFME}} + \mathcal{L}^{\mathrm{avg}}_{\mathrm{explore}}, \tag{52}$$

where $\mathcal{L}^{\mathrm{avg}}_{\mathrm{explore}}$ *is the average exploration loss along the trajectory.*

Proof. The vicinal risk over the explored trajectory is:

$$\mathcal{R}_{\mathrm{VRM}} = \frac{1}{M} \sum_{m=0}^{M-1} \ell(p_t(y \mid \mathbf{h}' + \epsilon^m), y)\,. \tag{53}$$

For the cross-entropy loss $\ell(p, y) = -\log p(y)$, we split the trajectory into two parts: (a) the original augmented feature $\mathbf{h}'$ (at $m = 0$, $\epsilon^0 = \mathbf{0}$), and (b) the final explored feature $\mathbf{h}' + \epsilon^M$ (at $m = M$). The BFME consistency loss measures the distributional discrepancy between these two endpoints:

$$\mathcal{L}_{\mathrm{BFME}} = \frac{1}{2}\Big(D_{KL}(A'_t \| \hat{A}_t) + D_{KL}(\hat{A}_t \| A'_t)\Big) = \mathrm{JSD}(A'_t \| \hat{A}_t). \tag{54}$$

By the log-sum inequality and convexity of cross-entropy:

$$\frac{1}{M} \sum_{m=0}^{M-1} \ell(p_t(y \mid \mathbf{h}' + \epsilon^m), y) \leq \frac{1}{2}\big(\ell(p_t(y \mid \mathbf{h}'), y) + \ell(p_t(y \mid \mathbf{h}' + \epsilon^M), y)\big) + \mathrm{JSD}(A'_t \| \hat{A}_t), \tag{55}$$

where the first two terms correspond to $\mathcal{L}^{\mathrm{avg}}_{\mathrm{explore}}$ evaluated at the two endpoints, and the JSD term accounts for the distributional deviation along the trajectory by convexity of KL divergence. Since $\mathrm{JSD}(A'_t \| \hat{A}_t) = \mathcal{L}_{\mathrm{BFME}}$:

$$\mathcal{R}_{\mathrm{VRM}} \leq \mathcal{L}_{\mathrm{BFME}} + \mathcal{L}^{\mathrm{avg}}_{\mathrm{explore}}. \tag{56}$$

This completes the proof. □

B Extended Related Work

Memory and optimization in continual learning. Beyond the FSCIL methods discussed in the main text, continual learning offers several ways to retain earlier knowledge De Lange et al. (2021); Parisi et al. (2019). Replay methods reuse past training information Rolnick et al. (2019); Buzzega et al. (2020), while regularization and gradient constraints limit interference during updates Zenke et al. (2017); Lopez-Paz & Ranzato (2017); Chaudhry et al. (2018). Auxiliary networks address the stability–plasticity trade-off Kim et al. (2023), and dual-view exploration has been studied in online class-incremental learning Gu et al. (2022).

Parameter-efficient adaptation. L2P, DualPrompt, and CODA-Prompt adapt pretrained models through lightweight prompts Wang et al. (2022b;a); Smith et al. (2023). Other work examines how pretrained representations and random projections support continual recognition Zhou et al. (2025); McDonnell et al. (2023). LoRA learns low-rank weight updates for a frozen pretrained model Hu et al. (2022). These studies provide context for choosing compact adaptation modules in incremental learning.

C Component Results and Reporting Conventions

C.1 Detailed Experimental Settings

Class split. Classes follow their original ID order on CIFAR-100/CUB-200 and alphabetical class-directory order on ImageNet-100. CUB-200 uses the first 100 classes for base training and divides the remaining 100 into ten consecutive groups of 10. CIFAR-100 and ImageNet-100 use the first 60 classes for base training and divide the remaining 40 into eight consecutive groups of five. Each group forms one incremental session.

Image split. We keep the official training and test sets of CIFAR-100/CUB-200; for ImageNet-100, the official `val` set serves as the test set. From the *training images of each base class*, we hold out 10 images on CIFAR-100, 3 on CUB-200, and 5 on ImageNet-100 to select the base checkpoint. The remaining base images train the base expert. For each novel class, five training images receive labels, and the remaining training images form the unlabeled pool. Test images are never used for training or model selection.

Random seeds. The base validation split uses seed 42 and stays fixed. The three runs (42, 1024, and 1993) vary optimization and novel-class support sampling, while keeping class groups and test sets unchanged. Each incremental run loads the base checkpoint trained with the same run seed.

Sampling and graph construction. Joint training assigns 25% of the sampling probability to labeled support when pseudo labels are available. Each graph-expansion round selects at most eight samples per class on CIFAR-100/ImageNet-100 and four on CUB-200. Graph propagation uses a closed-form solution on CIFAR-100 and 20 iterations on CUB-200/ImageNet-100.

C.2 Three-Seed Component Sensitivity on CIFAR-100

Table 2 reports the component summaries for seeds 42, 1024, and 1993. For the CIFAR-100 protocol, Avg. Global averages $\mathrm{Acc}^1, \ldots, \mathrm{Acc}^8$, whereas the main-table Avg. also includes Acc^0. In particular, $\mathrm{Avg.} = (\mathrm{Acc}^0 + 8\,\mathrm{Avg.\ Global})/9$. Accuracy differences in the main text are computed by subtracting the displayed means and are expressed in percentage points; RPD instead reports a relative percentage drop. Both drop measures compare cumulative evaluation sets as new classes are added, rather than tracking forgetting on a fixed set of old classes.

All four controlled variants reduce Final and Avg. Global accuracy, confirming that the three components make complementary contributions. Removing DSGE reduces these metrics by 3.07 and 3.06 points and increases RPD by 2.94 points. Replacing the evolving graph with a static graph produces larger losses of 4.15 and 4.05 points, showing that graph construction alone is insufficient when the representation and reliable sample set change over time. Removing ADES lowers both Final and Avg. Global by 3.49 points and increases RPD from 8.96 to 12.47. Interestingly, its Avg. New accuracy is 1.02 points higher than that of the full method, exposing a stability–plasticity trade-off: unconstrained adaptation can fit the current classes while degrading long-term cumulative performance. The largest degradation occurs without BFME, where Final and Avg. Global decrease

Variant	Final↑	Avg. Global↑	Avg. New↑	RPD↓
DSE-SSF	83.82±1.90	87.51±1.12	95.45±1.56	8.96±1.86
w/o DSGE	80.75±1.44	84.45±0.98	94.52±1.68	11.90±1.37
Static DSGE	79.67±2.00	83.46±1.14	94.53±1.55	13.08±1.96
w/o ADES	80.33±1.70	84.02±1.00	96.47±0.87	12.47±1.64
w/o BFME	78.56±2.29	82.42±1.23	93.42±1.60	14.21±2.27

Table 2: Core-component sensitivity on CIFAR-100 under the 60-base + 8-session protocol (%). Results are mean ± sample standard deviation over three seeds.

Dataset	Base epochs	Base time	Inc. sessions	Full incremental time
CIFAR-100	80	1.43–1.93 h	8	0.91–2.22 h
CUB-200	60	0.19–0.21 h	10	1.23–1.59 h
ImageNet-100	80	3.78–5.09 h	8	2.03–2.48 h

Table 3: End-to-end training cost per seed. Incremental time includes all sessions, evaluation, and checkpointing.

by 5.26 and 5.09 points, Avg. New decreases by 2.03 points, and RPD increases by 5.25 points. Given only three seeds, these results are interpreted through effect sizes and variability rather than formal significance claims.

C.3 Training Cost

All experiments use one CUDA device per run. Base training uses 80 epochs on CIFAR-100 and ImageNet-100 and 60 epochs on CUB-200. Each incremental session contains eight warm-up epochs and 30 joint-training epochs with four updates per epoch; router calibration uses cached features and does not require backbone back-propagation.

With rank-8 LoRA, only the current expert is optimized in each incremental session: 1.182M parameters for a 5-way session and 1.185M for a 10-way session. ADES temporarily maintains one equally sized gradient-free EMA expert, whereas DSGE and BFME add no trainable parameters. Their main overheads are the periodically refreshed k-NN graph and three classifier-level feature-space ascent steps, respectively.

C.4 Expanded Motivation Diagnostics

We analyze the three diagnostics summarized in Fig. 1 on the reproduced SSFSCIL-DTE baseline. The pseudo-label and adaptation analyses use the CIFAR-100 5-shot incremental setting, while the boundary analysis uses 1-, 2-, and 5-shot 5-way episodes. All summaries use three seeds; hidden labels are used only for post-hoc diagnosis or explicitly marked oracle controls. Figure 4 qualitatively positions DSE-SSF around reliable semantic discovery, incremental stability, and explicit boundary refinement. Experiments 1–3 then provide corresponding diagnostic evidence on selected supervision, old-class geometry, and few-shot decision boundaries, respectively, as context for the full-framework comparison in Fig. 3.

Static pseudo-labels contain persistent semantic errors. The pseudo-label set used by DTE has a $17.92\% \pm 0.31\%$ error rate. The highest observed confidence bin reaches 83.43% accuracy (16.6% errors) and contains 9.3 samples per session on average (Fig. 6(a)). The six most frequent directional class pairs account for 50.01% of all errors (Fig. 5). Thus, the observed errors are concentrated in a small set of recurring class confusions. This pattern motivates combining relations among samples with prediction confidence when constructing pseudo supervision. These errors also persist during training.

Naive prototype refresh ends with an 18.67% label error and yields only a 0.28-point Final improvement, whereas oracle correction improves Final by 1.45 points (Fig. 6(b,c)). Table 4 compares static selection, prototype refresh, oracle removal of incorrect pseudo labels, oracle correction, and training without pseudo supervision. These results show that confidence-based selection admits

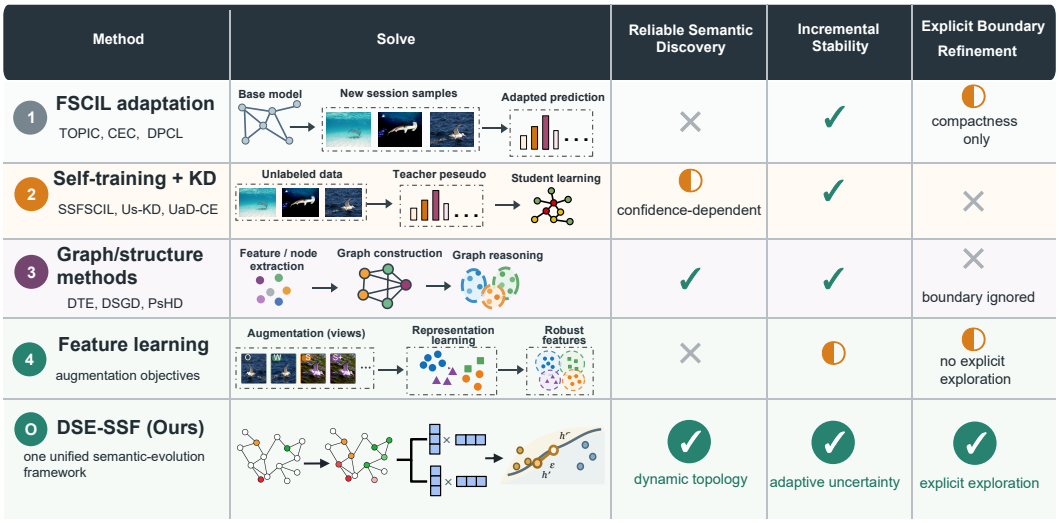

Figure 4: The differences between our approach and existing methods.

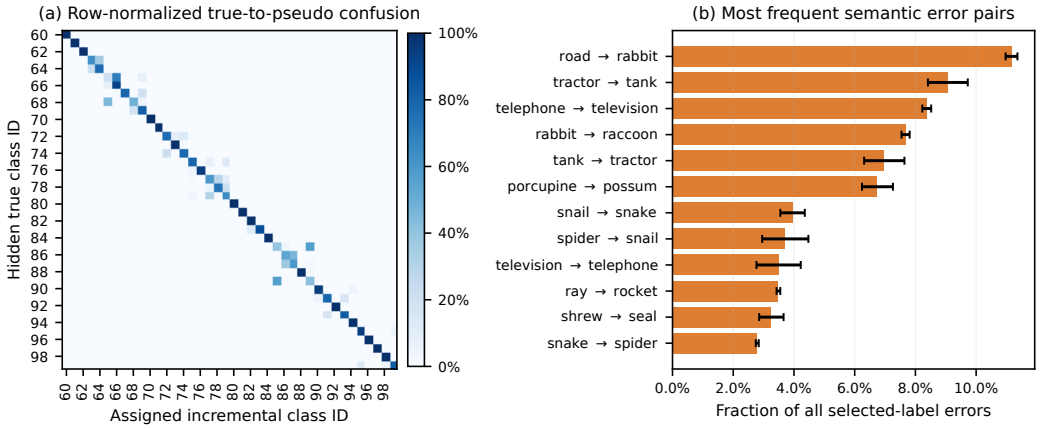

Figure 5: Semantic structure of static pseudo-label errors. (a) True-to-pseudo confusion and (b) the most frequent directional error pairs. Error bars show the standard deviation over three seeds.

systematic semantic errors with a measurable performance cost, while simple relabeling does not resolve them. This motivates DSGE to update pseudo supervision through evolving neighborhood structure and repeated reliability estimation.

Incremental adaptation reorganizes old-class geometry. We use a fixed old-class probe and a frozen-projection control to measure whether incremental adaptation changes local relationships among old classes. Relative to the frozen control, the fraction of fixed old-class samples that acquire new cross-class neighbors peaks at 9.49% and remains at $7.38\% \pm 0.21\%$ after training (Fig. 7(a)). In contrast, fixed-probe accuracy changes by at most 0.25 points. This contrast shows that local neighborhood membership can reveal adaptation-induced geometric changes that are not captured by fixed-probe accuracy alone.

Across the available session summaries within each seed, larger neighborhood drift is associated with a larger all-seen accuracy loss, with a mean within-seed Spearman correlation of -0.52 ± 0.17 (Fig. 7(b)). Thus, the measured neighborhood changes are related to cumulative recognition in this diagnostic rather than being reflected only as a change in feature coordinates. Together, these observations motivate treating old-class semantic geometry as an explicit stability target during incremental adaptation, providing the motivation for ADES.

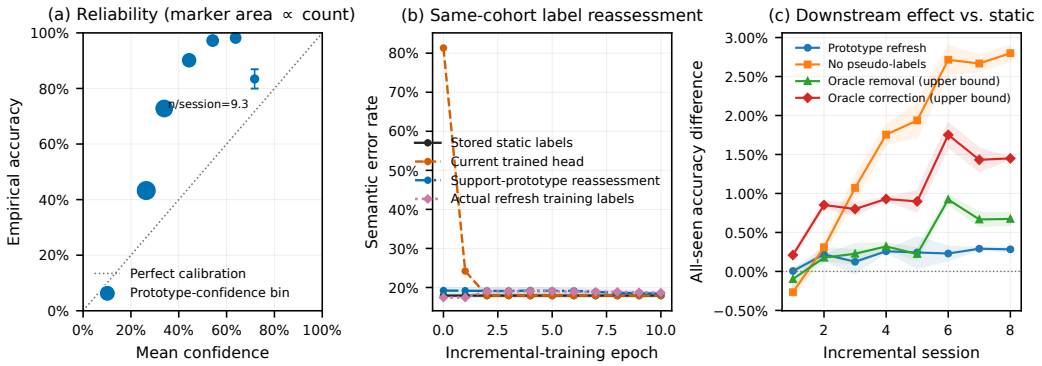

Figure 6: Controls for static pseudo-labeling. (a) Confidence reliability, (b) label reassessment during training, and (c) the downstream effect of pseudo-label interventions relative to static training.

Intervention	Final	Difference	Avg. Global
Static	36.74 ± 0.09	$+0.00$	46.78
Prototype refresh	37.02 ± 0.07	$+0.28$	46.99
Oracle removal	37.41 ± 0.13	$+0.67$	47.17
Oracle correction	38.19 ± 0.08	$+1.45$	47.82
No pseudo supervision	39.54 ± 0.01	$+2.80$	48.40

Table 4: All pseudo supervision interventions in the baseline diagnostic. Final is the last incremental-session accuracy (%); $\pm$ gives the reported seed standard deviation. The difference from Static is in percentage points. Avg. Global is computed from the eight incremental session means, excluding the base session.

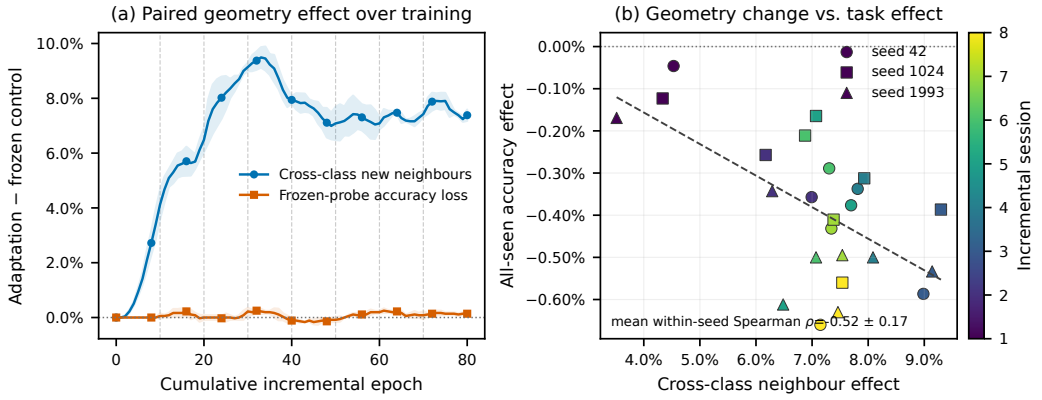

Figure 7: Representation drift relative to a frozen-projection control. (a) Paired geometry change over incremental training and (b) its association with the all-seen accuracy effect.

Few-shot boundaries remain support- and perturbation-sensitive. We use resampled 5-way episodes with cosine-prototype classifiers and vary the support size over 1, 2, and 5 shots to examine how support choice and feature perturbations affect prediction. As shown in Fig. 8(a–b), increasing the support size raises both clean accuracy and feature-perturbation accuracy at $\epsilon = 0.005$. For 5-shot episodes, accuracy changes from 62.20% without perturbation to 34.03% under the tested perturbation. The benefit of additional support also depends on the perturbation radius: at $\epsilon = 0.02$, feature-perturbation accuracy is below 0.5% for all three support sizes. Additional support therefore improves prediction most clearly at smaller tested radii.

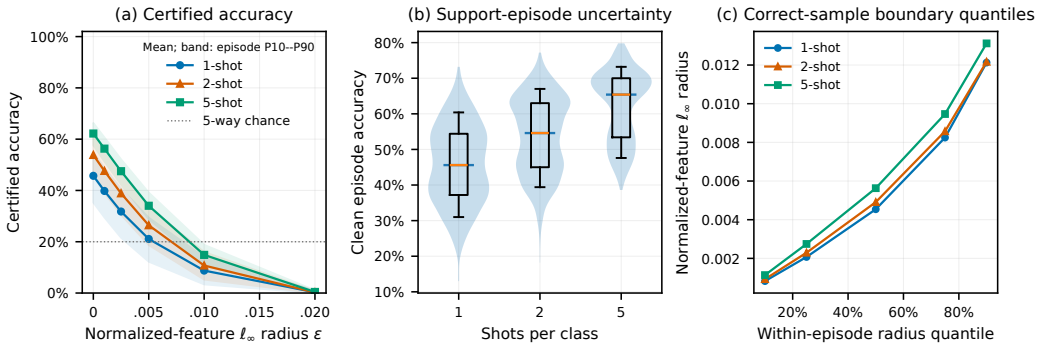

Figure 8: Few-shot boundary diagnostics over resampled 5-way episodes. (a) Certified accuracy under feature perturbations, (b) support-induced episode variability, and (c) local boundary-radius quantiles.

Support-induced variability decreases from 1 to 5 shots in both the mean within-session standard deviation and the accuracy range (Fig. 8(b)), with a mean range of 18.39 points in the 5-shot setting. The mean reported P10 boundary radius also increases with support size, reaching 0.001136 for 5-shot episodes (Fig. 8(c)). Together, these observations show that additional support improves accuracy and reduces variability, while support choice and proximity to decision boundaries remain relevant to prediction. This baseline analysis motivates BFME's exploration of feature regions beyond the observed support samples; the corresponding full-framework comparison is presented in Fig. 3(c).

Overall, the supplementary experiments establish a coherent motivation chain: DSGE addresses structured pseudo-label errors, ADES controls performance-linked representation drift, and BFME targets support-dependent low-margin boundaries. These three-seed CIFAR-100 diagnostics explain the observed baseline failures; the component results in Table 2 separately verify that the corresponding modules improve DSE-SSF.