

Flexi-FSCIL: Adaptive Knowledge Retention for Breaking the Stability-Plasticity Dilemma in Few-Shot Class-Incremental Learning

Wufei Xie¹ Yalin Wang¹ Chenliang Liu^{1†} Zhaohui Jiang¹ Xue Yang²

¹ Central South University, ² Shanghai Jiao Tong University

[†] Corresponding author lcliang@csu.edu.cn

Abstract

Few-Shot Class-Incremental Learning (FSCIL) is challenged by limited data and expanding class spaces, leading to overfitting and catastrophic forgetting. Existing methods, which often freeze feature extractors and use Nearest Class Mean classifiers, sacrifice adaptability to new feature distributions. To address these issues, we propose Flexi-FSCIL, a semi-supervised framework that integrates three novel strategies: Adaptive Gated Residual Fusion (AGRF), Attention-Guided Dynamic Hybrid Distillation (ADHD), and Prototype Offset Equilibrium (POE). Flexi-FSCIL effectively balances stability and plasticity in FSCIL. AGRF resolves the rigidity of frozen feature extractors by integrating both frozen and trainable components, enabling adaptive feature learning while retaining old-class knowledge. ADHD tackles the imbalance between old and new tasks by dynamically aligning features using cross-attention maps and direct matching, preserving old-class knowledge while facilitating new-class learning. POE addresses the issue of prototype drift in semi-supervised settings by selecting high-quality unlabeled samples, maintaining feature space separability and preventing overfitting. Evaluated on three benchmark datasets, Flexi-FSCIL achieves state-of-the-art performance, significantly outperforming existing FSCIL methods with only 12.97 performance drop on CUB200.

1. Introduction

With the rapid advancement of deep learning, deep neural networks have been widely applied to image classification [10], object detection [11], and segmentation [23]. However, real-world applications face two key challenges: (1) Few-shot data: Limited datasets often lead to overfitting, necessitating methods that generalize well in low-data regimes. (2) Class incrementality: The number of categories expands over time, making full re-training impractical and fine-tuning prone to catastrophic forgetting. These challenges have driven growing interest in Few-Shot Class-Incremental Learning (FSCIL).

To mitigate both catastrophic forgetting and overfitting, existing FSCIL methods often freeze the feature extractor and employ a Nearest Class Mean (NCM) classifier. Freezing the feature extractor ensures that previously learned representations remain unchanged, preventing knowledge degradation, while NCM avoids data-dependent updates, reducing overfitting. However, this rigid approach limits the ability to adapt to new feature distributions, reducing feature representation, which leads to spatial overlap between new and old classes and lowers generalization capacity in incremental learning. To address this, we propose Adaptive Gated Residual Fusion (AGRF), integrating frozen and trainable components to balance retention and adaptation, with a learnable gating mechanism for parameter control.

Knowledge distillation is also commonly used to mitigate forgetting, yet FSCIL methods that rely on frozen feature extractors cannot perform intermediate feature alignment. Most distillation techniques focus on aligning final output logits, assuming that the frozen feature extractor remains stable. However, intermediate features encode rich semantic information, and their misalignment can lead to performance degradation. Since AGRF incorporates trainable components, it allows for adaptive feature alignment. To better preserve old-class knowledge, we propose Attention-Guided Dynamic Hybrid Distillation (ADHD), which complements AGRF. Unlike rigid intermediate feature alignment, ADHD adopts a more flexible approach by first leveraging bidirectional cross-attention maps to guide feature alignment. To maintain feature space stability, we introduce a novel regularization mechanism. When attention maps exhibit significant discrepancies, making alignment ineffective, we dynamically incorporate direct alignment based on similarity measures.

However, introducing trainable fine-tuning components reintroduces overfitting in few-shot conditions. Leveraging unlabeled data for auxiliary training is an effective solution, making semi-supervised FSCIL (Semi-FSCIL) a promising direction. However, existing approaches overlook the impact of sample selection, which is crucial to avoid prototype

drift and maintain inter-class separability. To address this, we propose Prototype Offset Equilibrium (POE), enforcing feature affinity and separability constraints.

We integrate the AGRF and ADHD strategies into a Semi-FSCIL framework, proposing Flexi-FSCIL, a method that balances stability and plasticity, enabling effective knowledge retention and adaptation. Specifically, AGRF combines frozen and fine-tuned branches, while ADHD dynamically distills knowledge, preserving old-class memory while facilitating new-class learning. Additionally, POE ensures that only high-quality unlabeled samples are incorporated to prevent overfitting. Our main contributions are summarized as follows:

- We propose Adaptive Gated Residual Fusion (AGRF), a novel incremental training strategy that incorporates both frozen and trainable backbones to balance knowledge retention and adaptation. A learnable gating mechanism is introduced to control the ratio of old and new parameters, ensuring stable incremental learning.
- We introduce Attention-Guided Dynamic Hybrid Distillation (ADHD), a novel knowledge distillation strategy that employs cross-attention maps for soft feature alignment while dynamically integrating direct matching, thereby mitigating imbalance between old and new tasks.
- We propose a Prototype Offset Equilibrium (POE) strategy for semi-supervised FSCIL, ensuring that only high-quality unlabeled data is selected to prevent prototype drift and maintain feature space separability.
- We conduct extensive evaluations on three benchmark datasets to validate the effectiveness of our Flexi-FSCIL framework. The results demonstrate that our approach achieves SOTA performance and can further enhance existing state-of-the-art models when integrated.

2. Related works

2.1. Few-Shot Learning (FSL)

FSL enables models to adapt to new categories using only a handful of examples. Traditional FSL approaches fall into three categories: data augmentation, transfer learning, and meta-learning. Data augmentation methods [45] expand the dataset at the instance or feature level. Transfer learning [19] first pretrains a model on a large dataset and then fine-tunes it with techniques to mitigate overfitting. Meta-learning methods mainly follow three directions: metric-based [15, 27, 32], which measure similarity between support and query samples; optimization-based [9, 20], which optimize model parameters for rapid adaptation; and memory-based [25], which enhance learning by associating queries with stored knowledge.

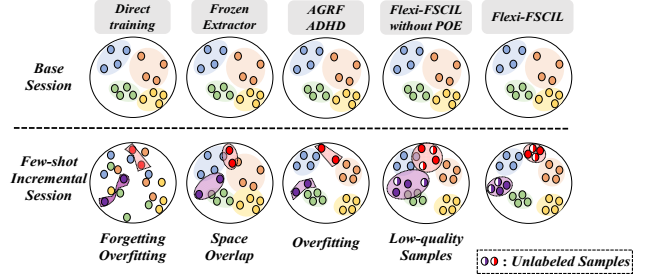


Figure 1. The motivation of our framework. Direct training suffers from forgetting as well as overfitting. Frozen extractors lead to overlapping feature spaces. And just adding AGRF with ADHD cannot solve the overfitting problem. Introducing semi-supervised learning directly without data filtering leads to low sample quality.

2.2. Class Incremental Learning (CIL)

CIL enables models to acquire new knowledge while preserving past information continuously. Traditional approaches fall into three categories: rehearsal-based methods [21, 41], which selectively retain past samples to maintain discrimination; regularization-based methods [2, 38], which prevent significant changes to important parameters; and architecture-based methods [35, 36], which expand the network to adapt to new tasks. However, CIL relies on large training datasets, making it unsuitable for Few-Shot Class-Incremental Learning (FSCIL). Directly applying CIL to the limited samples leads to overfitting.

2.3. Few-Shot Class-Incremental Learning (FSCIL)

FSCIL differs from conventional CIL by requiring knowledge retention while learning novel categories from limited data. Its main challenges include catastrophic forgetting in base class representations and poor generalization to new categories. Existing methods address these through spatial reservation and geometric constraints.

Spatial reservation techniques use synthetic category proxies to pre-allocate feature space. FACT [43] compresses base class embeddings via virtual prototypes, balancing knowledge retention and adaptability. SAVC [29] extends this with augmented contrastive learning, using virtual classes to enhance inter-class separation and serve as semantic anchors.

Geometric constraints optimize class distribution for future updates. NC-FSCIL [37] employs ETF-structured classifiers to maximize angular separation, while OrCo [1] introduces prototype perturbation and orthonormal projections to create buffer zones between classes.

Recent studies extend FSCIL to semi-supervised settings. Initial work [4] formalized the Semi-FSCIL paradigm, followed by uncertainty modeling [5, 6] for better knowledge integration. However, sample selection for training effects remains unexplored.

2.4. Semi-Supervised Learning (SSL)

SSL enhances supervised learning by leveraging unlabeled data. Existing SSL approaches can be broadly categorized into consistency regularization and pseudo-labeling. Consistency-regularization methods [3, 28, 39] enforce prediction stability under different perturbations, often relying on extensive data augmentation, which becomes less effective in low-data scenarios. Pseudo-labeling methods [12, 14, 16, 34] assign labels to unlabeled data using a model trained on labeled samples. These labels can be derived from neural network predictions [14, 34] or estimated via neighborhood graphs [12, 16].

3. Methodology

3.1. Preliminaries

We consider a sequence of tasks t utilizing disjoint datasets $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n\}$. The base dataset \mathcal{D}_1 contains sufficient data for training in the initial session, while subsequent datasets \mathcal{D}_i (for $i > 1$) are employed for N -way K -shot classification tasks in an incremental learning setting. Formally, the base dataset \mathcal{D}_1 is defined as: $\mathcal{D}_1 = \{(x_j, y_j)\}_{j=1}^{|\mathcal{D}_1|}$, where $y_j \in \mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{base}}$ represents the set of base classes. In the i -th session (where $i > 1$), the novel dataset \mathcal{D}_i is composed of labeled and unlabeled data, expressed as: $\mathcal{D}_i = \mathcal{D}_i^L \cup \mathcal{D}_i^U$. Here, $\mathcal{D}_i^L = \{(x_j, y_j)\}_{j=1}^{N \times K}$ denotes the labeled data, and $\mathcal{D}_i^U = \{x_j\}_{j=1}^{|\mathcal{D}_i^U|}$ represents the unlabeled data, with $|\mathcal{D}_i^U| \gg K$. Crucially, the classes across different sessions are non-overlapping, i.e., $\mathcal{C}_i \cap \mathcal{C}_{i'} = \emptyset$ for $i \neq i'$, ensuring that the model incrementally learns new classes without revisiting previously learned ones. The model $\mathcal{M}(\cdot)$ usually consists of a backbone $f_\theta(\cdot)$ parameterized by θ for feature extracting and the linear classifier $h_\phi(\cdot)$ parameterized by ϕ .

3.2. Overview of Flexi-FSCIL Framework

The overall architecture of our proposed Flexi-FSCIL framework is illustrated in Figure 2. In the initial base session, the model M_1 is trained on the base dataset D_1 , and a subset of D_1 is selected to construct the initial exemplar set ε_1 . In each subsequent incremental session, to better preserve the knowledge of previous models and mitigate catastrophic forgetting, the model M_i is derived from the previous model M_{i-1} by training on the novel dataset D_i using Adaptive Gated Residual Fusion (AGRF) strategy, with only a subset of parameters being updated. And knowledge from M_{i-1} is distilled into M_i through Attention-Guided Dynamic Hybrid Distillation (ADHD) strategy with the assistance of the exemplar set ε_{i-1} . The dataset D_i includes an unlabeled portion D_i^U , which is utilized for self-training to augment data for new classes, thereby alleviating overfitting caused by limited labeled samples. To enhance the

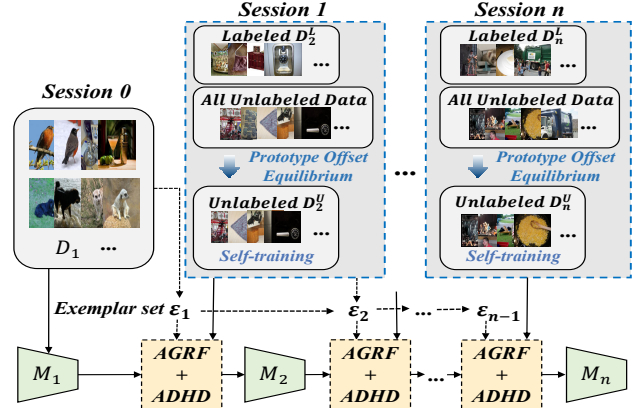


Figure 2. The overall of our Flexi-FSCIL framework. First, the model is trained on large-scale D_1 . Subsequently, numerous unlabeled samples are introduced and filtered via POE along with a few labeled samples. The filtered unlabeled and few-shot labeled samples are used to train the model using the AGRF strategy, while the exemplar set data is for the ADHD.

quality of unsupervised training, improve the utilization efficiency of unlabeled data, and mitigate the negative impact of noisy data, we adopt a Prototype Offset Equilibrium (POE) strategy to selectively filter high-quality unlabeled data for training. Upon completion of the current session, a subset of samples from the current dataset is selected and added to the exemplar set, updating it from ε_{i-1} to ε_i . Noteworthy, the exemplar set will not be updated in the final session.

3.3. Adaptive Gated Residual Fusion

The balance between old and new classes is the primary consideration. As demonstrated in [26], the performance of incremental freezing surpasses that of many trainable methods, indicating that preserving knowledge of old classes may take precedence over acquiring new class knowledge. Therefore, as illustrated in Figure 3, we adopt an Adaptive Gated Residual Fusion (AGRF) approach that leverages a freezing strategy to retain old class knowledge while incorporating a trainable residual branch to facilitate the learning of new class knowledge.

When integrating the frozen parameters of the old model with the trainable parameters of the new model, we refrain from directly summing the parameters of each network layer. This is because different layers exhibit varying sensitivities to old and new knowledge. A straightforward summation would impose equal-weighted overlap across all dimensions, leading to contamination of lower-level general feature parameters by the new parameters. In contrast, high-level semantic parameters would struggle to adapt to the new classes fully.

To better balance the retention of old class knowledge and the acquisition of new class knowledge, we introduce a learnable evaluation parameter, constructing a bidirectional

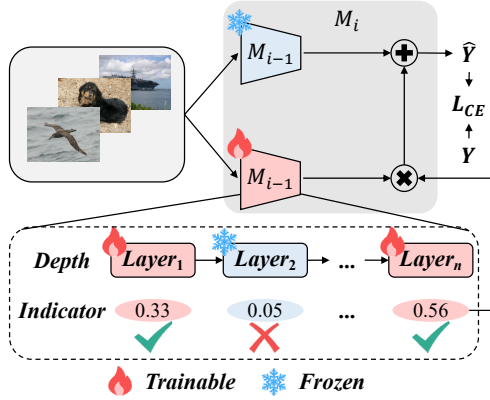


Figure 3. The illustration of our Adaptive Gated Residual Fusion (AGRF) strategy. The model combines a frozen branch and a trainable residual branch, where learnable coefficients control the hierarchical training and update the parameters by multiplying the training parameters by the corresponding coefficients in combination with the frozen parameters.

gating mechanism in parameter space to guide the fusion of the two branches. The fused parameters at the i th layer and the final outputs during training can be expressed as:

$$\begin{aligned} \theta^i &= (1 - G_\tau(W^i)) * \theta_{\text{frozen}}^i + G_\tau(W^i) * \theta_{\text{new}}^i, \\ \hat{y} &= (1 - G_\tau(W^i)) * \hat{y}_{\text{frozen}} + G_\tau(W^i) * \hat{y}_{\text{new}}, \end{aligned} \quad (1)$$

where W^i is a learnable evaluation parameter that assesses the necessity of parameter fusion at the i th layer and determines the corresponding parameter weights. $G_\tau(\cdot)$ is a threshold function designed to control whether parameter fusion should be performed, defined as:

$$G_\tau(W^i) = \begin{cases} W^i & \text{if } W^i \geq \tau \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where τ is a hyperparameter. Parameter fusion between the old and new model components at the corresponding layer is performed only when W^i exceeds the threshold τ .

3.4. Attention-Guided Dynamic Hybrid Distillation

Current incremental learning methods often preserve old-class knowledge via knowledge distillation, yet predominantly align output layers of old and new models [8, 42], neglecting fine-grained information in intermediate features (e.g., textures, spatial structures). This results in insufficient representation robustness and compromised retention of old-task feature distributions. While direct intermediate feature alignment addresses this limitation, rigid enforcement may over-constrain new task learning, causing the imbalance between old and new tasks.

To address this, we propose an Attention-Guided Dynamic Hybrid Distillation (ADHD) strategy, which employs indirect attention-guided structural knowledge transfer and

conditionally activates direct feature alignment based on semantic relevance. As illustrated in Figure 4, ADHD operates in two phases:

1. **Mutual Attention Interaction:** After extracting intermediate features F_{i-1}^k and F_i^k from the old model M_{i-1} and the incremental model M_i , we compute old→new attention $A_{o \rightarrow n}^k$ using F_{i-1}^k as Query and F_i^k as Key-Value, and vice versa for new→old attention $A_{n \rightarrow o}^k$. The loss that enforces their alignment is shown as follows:

$$L_{\text{attn}} = \sum_{k=0}^N (\|A_{o \rightarrow n}^k - A_{n \rightarrow o}^k\|_1 + \alpha \|A_{o \rightarrow n}^k A_{n \rightarrow o}^k - I\|_F^2), \quad (3)$$

where N is the layer number. The former term is a normal L_1 loss for aligning the attention graph, while the latter is an orthogonal constraint term for maintaining the structural stability of the feature space. α determines the extent to which orthogonal constraints are enforced

2. **Conditional Feature Alignment:** We first apply global average pooling to $A_{o \rightarrow n}^k$ and $A_{n \rightarrow o}^k$ to aggregate their information into compact feature vectors. The cosine similarity is then computed between these pooled representations. When the similarity is low, indicating that the indirect supervision provided by the attention alignment is insufficient, a direct feature alignment loss is dynamically activated:

$$\begin{aligned} A_1^k &= \text{AvgPool}(A_{o \rightarrow n}^k), \quad A_2^k = \text{AvgPool}(A_{n \rightarrow o}^k), \\ w_k &= \text{ReLU}(\beta - \text{sim}(A_1^k, A_2^k)), \\ L_f &= \sum_{k=0}^N w_k * \|F_{i-1}^k - F_i^k\|_1, \end{aligned} \quad (4)$$

where $\text{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y} / (\|\mathbf{x}\| \|\mathbf{y}\|)$ is cosine similarity between two vectors, β is an adaptive threshold. The ReLU function is introduced to ensure smooth transitions.

Thus, the overall loss for the entire model is as followed:

$$L = (1 - \lambda)L_{ce} + \lambda(L_{kl} + L_{\text{attn}} + L_f), \quad (5)$$

where L_{ce} represents the cross-entropy loss between the model predictions and the true labels, and L_{kl} represents the kl loss between the old and new model output. λ balances the L_{ce} with the loss for knowledge distillation.

3.5. Prototype Offset Equilibrium

To address the overfitting challenge in few-shot incremental session training, we introduce a semi-supervised learning framework that incorporates a large-scale unlabeled dataset for self-training. However, directly utilizing unlabeled data raises critical challenges: (1) Noisy pseudo-labels may cause prototype drifting in feature space; (2) Indiscriminate data selection could compromise inter-class separability. This necessitates a principled approach to select unlabeled samples that enhance class representation while preserving feature space discriminability.

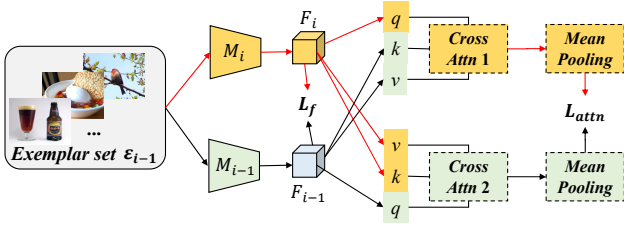


Figure 4. The illustration of Attention-Guided Dynamic Hybrid Distillation (ADHD) strategy. Soft alignment through cross-attention ensures the acquisition of new knowledge and the transfer of old knowledge.

Our Prototype Offset Equilibrium (POE) method addresses this dual objective through geometric constraints in the embedding space. The core insight is that qualified samples should satisfy two conditions: (a) Feature affinity (FA): The sample’s embedding must exhibit maximum similarity with its corresponding class prototype compared to all other classes; (b) Separation preservation (SP): Incorporating the sample should not disproportionately increase inter-class prototype similarity. Formally, let p_c denote the prototype for class c computed as the mean embedding of labeled samples. For an unlabeled sample x with pseudo-label y_u , we first calculate its normalized similarity distribution across all prototypes:

$$s_c(\mathbf{x}) = \frac{\exp(\text{sim}(f_\theta(x), p_c))}{\sum_{i=1}^C \exp(\text{sim}(f_\theta(x), p_i))}, \quad (6)$$

where C is the total class number. And for feature affinity condition, the selected samples are required:

$$s_{y_u}(\mathbf{x}) > \max_{c \neq y_u} s_c(\mathbf{x}). \quad (7)$$

To enforce separation preservation, we monitor the inter-class similarity shift before and after the prototype update:

$$\Delta_{\text{in}} = \frac{1}{C-1} \sum_{i \neq y_u} [\text{sim}(p'_{y_u}, p_i) - \text{sim}(p_{y_u}, p_i)] > \sigma, \quad (8)$$

where p_{y_u} represents the updated prototype. Samples causing Δ_{in} beyond threshold σ are rejected. This dual-criteria selection ensures the augmented prototypes maintain both intra-class compactness and inter-class distinctiveness.

4. Experiments

4.1. Experimental Details

Datasets and Evaluation Indicators Following established experimental protocols in continual learning research [29, 30, 43], we evaluate our approach on three benchmark datasets: 1. **CIFAR100** [13]: This dataset contains 100 object categories with 600 samples per class. Our implementation divides these into 60 base classes for initial training and 40 classes for incremental learning phases, following a 5-way 5-shot configuration for incremental sessions.

2. **miniImageNet** [24]: As a condensed version of ImageNet [7], this dataset comprises 100 carefully selected classes. Mirroring the CIFAR100 configuration, we allocate 60 classes for the base session and 40 classes for incremental updates, maintaining the 5-way 5-shot learning paradigm. 3. **CUB200** [33]: This fine-grained dataset features 200 bird species. We employ a different split strategy with 100 base classes and 100 incremental classes, adopting a 10-way 5-shot configuration for incremental phases to accommodate its larger class diversity.

We conducted on two evaluation indicators: 1) Δ_F : The final improvement in performance compared to the Finetune baseline [29] in the last session; and 2) PD: The difference between the first and last session.

Implementation Details We evaluate Flexi-FSCIL independently and integrate it into SOTA methods, i.e. FACT [43], SAVC [29], and FACL [17]. To ensure fairness, all methods use ResNet18 [10] for *miniImageNet* and CUB200, and ResNet20 [10] for CIFAR100. Training is performed using SGD [22] ($\text{lr} = 0.1$, $\text{wd} = 5e-4$). For exemplar selection, we adopt the herding-based approach [21]. The framework is implemented in PyTorch and trained on GeForce RTX 4090 GPUs.

For CIFAR100 and *miniImageNet*, the initial learning rate is 0.1, reduced by a factor of 10 at epochs 80 and 120, with a total of 160 epochs. In subsequent sessions, it is set to 0.001 for 100 epochs. For CUB200, the first-session learning rate starts at 0.001, decreasing at epochs 80 and 120, and remains 0.0005 for 60 epochs in later sessions. The batch size is 128 for *miniImageNet*, 32 for CIFAR100 and CUB200, with test batch sizes of 100 and 50, respectively. Hyper-parameter selection in Flexi-FSCIL is: $\tau = 0.1$, $\alpha = 0.1$, $\beta = 0.8$, $\lambda = 0.3$. The experiments about hyper-parameter are discussed in Appendix.

For CIFAR100, we introduce 350 unlabeled samples over 35 iterations, while *miniImageNet* and CUB200 incorporate 160 samples across 16 iterations. Each iteration selects 10 class-balanced samples. Training with pseudo-labels extends by 10 epochs per iteration for CIFAR100 and *miniImageNet*, and 20 epochs for CUB200.

4.2. Performance on Benchmarks

We compare our Flexi-FSCIL framework with three FSCIL benchmark methods, CIL methods iCaRL [21], FSCIL-specific approaches like FSCIL methods including SPPR [44], CEC [40], NC-FSCIL[37], CLOSER [18], FACT [43], SAVC [29], FACL [17], and Semi-FSCIL methods including SS-iCaRL [4], Semi-SPPR [6], Semi-CEC [6], Us-KD [5], UaD-CIE [6]. Additionally, we evaluate the performance of FACT [43], SAVC [29], FACL [17] embedded with Flexi-FSCIL. All the three methods adopt the frozen paradigm and NCM. We include results from a basic ‘finetune’ [29] strategy for comparative purposes. We present

Table 1. Performance comparison on CUB200 dataset.

Method	Accuracy in each session (%) \uparrow										$\Delta F \uparrow$	PD \downarrow	
	0	1	2	3	4	5	6	7	8	9			10
Finetune [29]	68.68	43.70	25.05	17.72	18.08	16.95	15.10	10.06	8.93	8.93	8.47	-	60.21
iCaRL [21]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	+12.69	47.52
SPPR [44]	68.68	61.85	57.43	52.68	50.19	46.88	44.65	43.07	40.17	39.63	37.33	+28.86	31.35
CEC [40]	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	+43.81	23.57
NC-FSCIL [37]	80.45	75.98	72.30	70.28	68.17	65.16	64.43	63.25	60.66	60.01	59.44	+50.97	21.01
CLOSER [18]	79.40	75.92	73.50	70.47	69.24	67.22	66.73	65.69	64.00	64.02	63.58	+55.11	15.82
FACT [43]	77.66	73.70	70.23	65.97	65.24	61.92	61.20	59.51	57.81	57.35	56.11	+47.64	21.55
SAVC [29]	82.21	78.30	75.25	70.65	70.11	66.96	66.36	65.32	63.61	63.86	62.14	+53.67	20.07
FACL [17]	82.74	80.09	76.89	71.31	70.51	68.12	67.54	66.97	66.05	65.39	64.70	+56.23	18.04
SS-iCaRL [4]	69.89	61.24	55.81	50.99	48.18	46.91	43.99	39.78	37.50	34.54	31.33	+22.86	38.56
Semi-SPPR [6]	68.44	61.66	57.11	53.41	50.15	46.68	44.93	43.21	40.61	39.21	37.43	+28.96	31.03
Semi-CEC [6]	75.82	71.91	68.52	63.53	62.45	58.27	57.62	55.81	54.85	53.52	52.26	+43.79	23.56
Us-KD [5]	74.69	71.71	69.04	65.08	63.60	60.96	59.06	58.68	57.01	56.41	55.54	+47.07	19.15
UaD-CIE [6]	75.17	73.27	70.87	67.14	65.49	63.66	62.42	62.55	60.99	60.48	60.72	+52.25	14.45
Flexi-FSCIL	79.21	77.58	74.26	71.90	70.25	68.92	68.15	67.80	66.84	66.65	66.24	+57.77	12.97

Table 2. Performance comparison on CIFAR100 dataset.

Method	Accuracy in each session (%) \uparrow										$\Delta F \uparrow$	PD \downarrow	
	0	1	2	3	4	5	6	7	8	9			10
Finetune [29]	64.10	39.61	15.37	9.80	6.67	3.80	3.70	3.14	2.65	-	-	-	61.45
iCaRL [21]	64.10	53.28	41.69	34.13	27.93	25.06	20.41	15.48	13.73	+11.08	50.37	-	50.37
SPPR [44]	63.97	65.86	61.31	57.60	53.39	50.93	48.27	45.36	43.32	+40.67	20.65	-	20.65
CEC [40]	73.07	68.88	65.26	61.19	58.09	55.57	53.22	51.34	49.14	+46.49	23.93	-	23.93
NC-FSCIL [37]	82.52	76.82	73.34	69.68	66.19	62.85	60.96	65.32	59.02	+56.37	26.41	-	26.41
CLOSER [18]	75.72	71.83	68.32	64.62	61.91	59.25	57.53	55.43	53.32	+50.67	22.40	-	22.40
FACT [43]	74.60	72.09	67.56	63.52	61.38	58.36	56.28	54.24	52.10	+49.45	22.50	-	22.50
SAVC [29]	80.93	75.77	71.96	68.44	65.71	62.22	59.12	57.18	55.98	+53.33	24.95	-	24.95
FACL [17]	86.20	81.55	76.95	72.50	68.75	65.68	63.16	60.62	58.20	+55.55	28.00	-	28.00
SS-iCaRL [4]	64.13	56.02	51.16	50.93	43.46	41.69	38.41	39.25	34.80	+32.15	29.33	-	29.33
Semi-SPPR [6]	76.68	72.63	67.59	63.69	59.24	56.02	53.23	50.46	48.29	+45.64	28.39	-	28.39
Semi-CEC [6]	73.03	70.72	65.79	61.91	58.64	55.84	53.70	51.37	49.37	+46.72	23.66	-	23.66
Us-KD [5]	76.85	69.87	65.46	62.36	59.86	57.29	55.22	54.91	54.42	+51.77	22.43	-	22.43
UaD-CIE [6]	75.55	72.17	68.57	65.35	62.80	60.27	59.12	57.05	54.50	+51.85	21.05	-	21.05
Flexi-FSCIL	79.54	76.35	73.89	70.07	68.51	66.43	63.17	61.33	59.78	+57.13	19.76	-	19.76

Table 3. Performance comparison on *miniImageNet* dataset

Method	Accuracy in each session (%) \uparrow										$\Delta F \uparrow$	PD \downarrow	
	0	1	2	3	4	5	6	7	8	9			10
Finetune [29]	61.31	27.22	16.37	6.08	2.54	1.56	1.93	2.60	1.40	-	-	-	59.91
iCaRL [21]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	+15.81	44.10	-	44.10
SPPR [44]	61.45	63.80	59.53	55.53	52.50	49.60	46.69	43.79	41.92	+40.52	19.53	-	19.53
CEC [40]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	+46.23	24.37	-	24.37
NC-FSCIL [37]	84.02	76.80	72.00	67.83	66.35	64.04	61.46	59.54	58.31	+56.91	25.71	-	25.71
CLOSER [18]	76.02	71.61	67.99	64.69	61.70	58.94	56.23	54.52	53.33	+51.93	22.69	-	22.69
FACT [43]	72.56	69.63	66.38	62.77	60.60	57.33	54.34	52.16	50.49	+49.09	22.07	-	22.07
SAVC [29]	81.12	76.14	72.43	68.92	66.48	62.95	59.92	58.39	57.11	+55.71	24.01	-	24.01
FACL [17]	86.68	81.49	76.65	72.65	69.71	66.02	63.08	61.17	59.48	+58.08	27.20	-	27.20
SS-iCaRL [4]	62.98	60.88	57.63	52.80	50.66	48.28	45.27	41.65	40.51	+39.11	22.47	-	22.47
Semi-SPPR [6]	80.10	74.21	69.31	64.83	60.53	57.36	53.70	52.01	49.61	+48.21	30.49	-	30.49
Semi-CEC [6]	71.91	66.81	63.87	59.41	56.42	53.83	51.92	49.57	47.58	+46.18	24.33	-	24.33
Us-KD [5]	72.35	67.22	62.41	59.85	57.81	55.52	52.64	50.86	50.47	+49.07	21.88	-	21.88
UaD-CIE [6]	72.35	66.91	62.13	59.89	57.41	55.52	53.26	51.46	50.52	+49.12	21.83	-	21.83
Flexi-FSCIL	77.37	74.96	70.73	66.50	63.95	62.44	61.49	60.29	59.69	+58.29	19.76	-	19.76

the performance results of our evaluation over the benchmark datasets: CUB200, CIFAR100, and *miniImageNet* in Table 1, 2, 3.

Except for embedding Flexi-FSCIL in other methods, on the CUB200 dataset, as demonstrated in Table 1, our method outperformed all others, achieving a 57.77% ΔF and a 12.97% in PD. Compared with other methods, our approach achieved superior ΔF and PD improvements of 1.54% and 1.48%, respectively. On the CIFAR100 dataset, we observed similar results, with a 1.55% increase in ΔF and a 1.29% improvement in PD, as shown in Table 2. On the *miniImageNet* dataset, our method performed best, with a 0.21% ΔF improvement and a 4.15% PD improvement, detailed results are provided in Table 3. Notably, on the *miniImageNet* dataset, the enhancement in ΔF is less pronounced compared to other datasets. This can be attributed to FACL’s superior performance in base sessions - despite its higher performance degradation rate, it maintains relatively high final accuracy. Our method prioritizes incremental class training over optimizing base classes to achieve knowledge acquisition without catastrophic forgetting while maintaining the balance between old and incremental classes. Overall, our approach outperforms the current state-of-the-art methods across all datasets.

We also explored the compatibility between our methods and the frozen methods, and the results are shown in Table 4. When is embedded into FACT, SAVC, and FACL, the methods have shown promising compatibility, demonstrating exceptional generalization capabilities when com-

pared with freezing strategies. By integrating our approach into these existing methodologies, we achieve a more flexible training process with trainable parts, rather than relying solely on freezing techniques. The PD metric improvements across three datasets stand out, recording increments of 1.23%, 0.63% and 1.20% on CUB200 dataset respectively. The improvements of ΔF and PD are the same for no accuracy change in the base session. These results underscore the potential of our strategy when introduced into conventional approaches, effectively mitigating forgetting rates while avoiding excessive memory loss compared to freezing-only methods. This demonstration highlights the scalable applicability and promising future of our method in conjunction with various complementary strategies.

4.3. Ablation study

Effectiveness of strategies: To evaluate the effectiveness of our proposed strategies, we conducted experiments on the CUB200 dataset and incorporated techniques such as AGRF, ADHD, and POE. Additionally, to highlight the impact of POE, we included a semi-supervised approach that does not utilize POE as a comparison. As shown in Table 5, methods without any optimization strategies demonstrated significant knowledge retention issues during subsequent incremental conversations and failed to maintain effective knowledge storage. Each individual strategy demonstrated improvements in performance metrics when implemented alone. The AGRF strategy significantly enhanced the system’s ability to remember knowledge while also leading to

Table 4. Comparative experimental results of freezing models w/o Flexi-FSCIL embedding on CUB200 dataset

Method	Accuracy in each session (%) \uparrow										$\Delta_F \uparrow$	PD \downarrow	
	0	1	2	3	4	5	6	7	8	9			10
Finetune [29]	68.68	43.70	25.05	17.72	18.08	16.95	15.10	10.06	8.93	8.93	8.47	-	60.21
FACT [43]	77.66	73.70	70.23	65.97	65.24	61.92	61.20	59.51	57.81	57.35	56.11	+47.64	21.55
SAVC [29]	82.21	78.30	75.25	70.65	70.11	66.96	66.36	65.32	63.61	63.86	62.14	+53.67	20.07
FACL [17]	82.74	80.09	76.89	71.31	70.51	68.12	67.54	66.97	66.05	65.39	64.70	+56.23	18.04
FACT+Flexi	77.66	74.65	71.08	66.39	65.81	63.16	61.99	60.30	58.69	58.02	57.34	+48.87	20.32
SAVC+Flexi	82.21	78.81	75.89	71.93	71.13	68.15	66.86	66.07	64.15	63.97	62.77	+54.30	19.44
FACL+Flexi	82.74	80.63	77.32	73.98	71.67	70.03	68.82	68.31	67.41	66.55	65.90	+57.43	16.84
Flexi-FSCIL	79.21	77.58	74.26	71.90	70.25	68.92	68.15	67.80	66.84	66.65	66.24	+57.77	12.97

Table 5. Ablation studies on the CUB200 benchmark for effectiveness of strategies

AGRF	ADHD	Unlabel	POE	Accuracy in each session (%)										Δ_{Acc}	
				0	1	2	3	4	5	6	7	8	9		10
				79.21	48.95	35.67	30.19	28.09	26.85	25.07	22.43	19.69	18.71	18.69	-
✓				79.21	73.11	66.29	63.78	60.54	58.73	56.19	54.98	53.03	52.86	52.77	+34.08
	✓			79.21	76.02	68.83	65.85	61.57	59.23	58.88	57.60	56.33	55.92	55.39	+36.70
		✓		79.21	77.06	74.26	69.71	64.79	61.42	59.37	59.17	58.55	58.48	58.32	+39.63
✓	✓	✓	✓	79.21	77.58	74.26	71.90	70.25	68.92	68.15	67.80	66.84	66.65	66.24	+47.55

substantial improvements in overall performance. Furthermore, the ‘semi-freezing’ optimization strategy effectively prevented catastrophic forgetting by helping the network avoid large-scale class disappearance events and maintaining critical functional capabilities. When only a semi-supervised approach without POE was employed, significant performance gains were observed in the initial incremental conversation tasks; however, these benefits were not sustained in later stages. The inclusion of POE eliminated such late-stage performance degradation. This observation may be attributed to the increased number of classes and tighter inter-class boundaries during later testing stages, necessitating effective filtering techniques to fully leverage the added unlabeled data.

Visualization: In this study, we utilized the t-SNE algorithm [31] to visualize the feature space. For this analysis, we randomly selected six classes from the initial conversations (0-5) and three classes from the incremental conversations (6-8). As shown in Figure 5, without any strategy employed, the feature spaces of the classes exhibited significant overlap, with all classes being conflated together. In contrast, when incorporating the AGRF and ADHD strategies, the feature space demonstrated a noticeable degree of separation, particularly for the initial classes. However, due to overfitting issues, subsequent new classes were unable to achieve satisfactory separation. When only semi-supervised methods were added without POE, the incremental classes achieved some level of separation in their feature spaces, but this may have been limited by the selected unlabeled data, causing prototype distances to remain relatively close and resulting in incremental classes still being intertwined with initial classes in the feature space. After the integration of the POE strategy, the model successfully separated all classes in their feature spaces.

Effectiveness of POE selection metrics: We also conduct some experiments to validate the effectiveness of our POE selection metrics. As shown in Table 7, experimental re-

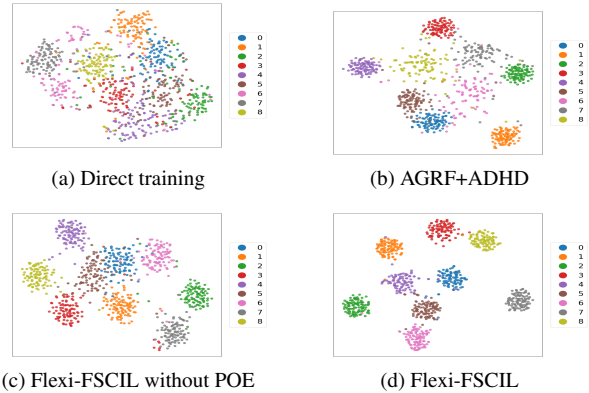


Figure 5. Comparison of t-SNE plots for ablation study on the CUB200 dataset (best view in color).

sults indicate that both FA and SP strategies contribute to unlabeled data selection and performance improvement in incremental session training, each playing a distinct role. When compared to the baseline, the inclusion of either FA or SP enhances performance. With FA alone, initial performance is poorer due to the potential introduction of noisy samples, causing instability in class prototypes. However, as training progresses, FA strengthens prototype representativeness, leading to significant performance gains in later sessions. In contrast, SP maintains inter-class separability in the early stages, resulting in better early performance. However, SP does not enhance intra-class consistency, causing performance improvement to plateau over time. When both FA and SP are applied simultaneously, the model achieves the best results across all sessions. FA improves intra-class compactness and representativeness, while SP preserves inter-class separability, especially benefiting larger class spaces. FA is more effective in scenarios with many classes, whereas SP excels in fewer-class settings. When combined, FA and SP complement each other, with FA enhancing class compactness and SP ensuring sep-

Table 6. Ablation studies on the CUB200 benchmark for number of unlabeled sample

Unlabeled Num		Accuracy in each session (%)										Δ_{Acc}	
		0	1	2	3	4	5	6	7	8	9		10
0	-	79.21	74.02	68.83	65.85	61.57	59.23	58.88	57.60	56.33	55.92	55.39	-
40	w/o POE	79.21	74.98	70.73	67.51	62.81	60.74	59.06	58.34	57.25	56.89	56.93	+1.18
	POE	79.21	76.61	73.62	69.54	68.16	66.83	64.92	64.38	63.83	63.47	63.19	+7.8
80	w/o POE	79.21	75.34	71.85	68.39	64.33	61.98	59.49	58.87	57.33	57.05	56.93	+1.54
	POE	79.21	76.93	74.25	70.83	68.95	66.71	65.32	64.77	64.51	64.29	64.17	+8.78
120	w/o POE	79.21	77.06	74.26	69.71	66.79	64.42	62.37	62.17	61.55	60.48	60.32	+4.93
	POE	79.21	77.51	73.87	71.63	68.19	67.18	66.49	66.00	65.54	65.37	65.02	+9.63
160	w/o POE	79.21	75.65	73.16	69.27	66.23	63.11	62.48	61.75	60.06	59.87	59.51	+4.12
	POE	79.21	77.58	74.26	71.90	70.25	68.92	68.15	67.80	66.84	66.65	66.24	+10.85
200	w/o POE	79.21	76.89	73.15	68.57	65.76	62.14	60.45	59.06	57.33	57.01	56.96	+1.57
	POE	79.21	77.44	74.17	71.29	68.64	66.98	65.74	64.95	64.69	64.31	64.02	+8.63

Table 7. Ablation studies on the CUB200 benchmark for effectiveness of POE selection metrics.

FA	SP	Accuracy in each session (%)										Δ_{Acc}	
		0	1	2	3	4	5	6	7	8	9		10
		79.21	75.65	73.16	69.27	66.23	63.11	62.48	61.75	60.06	59.87	59.51	-
	✓	79.21	76.59	73.60	71.44	69.92	68.54	67.80	67.62	66.75	65.77	65.59	+6.08
	✓	79.21	77.25	74.08	71.58	68.96	67.63	67.36	67.63	65.31	64.67	64.36	+4.85
	✓	79.21	77.58	74.26	71.90	70.25	68.92	68.15	67.80	66.84	66.65	66.24	+6.73

Table 8. Ablation studies on the CUB200 benchmark for the calculation way of similarity

sim($A_{o \rightarrow n}, A_{n \rightarrow o}$)	Accuracy in each session (%)										PD	
	0	1	2	3	4	5	6	7	8	9		10
By Spital	79.21	76.67	71.86	70.28	68.93	67.92	67.25	66.48	65.98	65.92	65.12	14.09
By Channel	79.21	76.87	72.72	71.12	69.12	68.59	67.59	66.64	66.11	65.55	65.01	14.20
Global Flatten	79.21	77.13	73.49	71.33	70.02	68.85	67.97	67.44	66.78	65.81	65.97	13.24
AvgPool	79.21	77.58	74.26	71.90	70.25	68.92	68.15	67.80	66.84	66.65	66.24	12.97

arability, ultimately leading to optimal performance.

Number of unlabeled samples: In this study, we conducted experiments on the CUB200 dataset to investigate the optimal quantity of unlabeled data. The results, as shown in Table 6, reveal that when the number of unlabeled samples is relatively small, an increased proportion of unlabeled samples leads to more pronounced performance improvements, effectively mitigating overfitting issues by using a large amount of labeled data for model training. However, the quantity of samples is not universally advantageous; when the sample size becomes excessive, it might lead to performance degradation due to noise introduced by unlabeled data. Furthermore, the results demonstrate that the incorporation of the POE method enables models to utilize a greater volume of unlabeled data for training without compromising performance. This allows for a more effective exploitation of semi-supervised learning capabilities, balancing the potential benefits and drawbacks associated with different quantities of unlabeled samples.

Calculation way of similarity: After obtaining the attention vectors, various methods can be employed to calculate their cosine similarity, each impacting performance differently. We compared four approaches: spatial average, channel-wise average, global flatten, and global average pooling. The experimental results, as shown in Table 8, indicate that neither spatial nor channel-wise yields good performance. The spatial average overly focuses on local details at each position, making it susceptible to noise and minor positional shifts, leading to inconsistencies in lo-

cal activations. Channel-wise average, on the other hand, overlooks the holistic semantic information across spatial regions; the high-dimensional channel data may amplify redundancy and noise, adversely affecting alignment. While the global average after flatten retains all information and better integrates spatial and channel data, resulting in improved performance, the high-dimensional data still incorporates substantial local noise and redundant features, rendering similarity computations less robust. In contrast, global average pooling effectively suppresses local noise during aggregation, preserving only the global statistical information of each channel, resulting in more stable and accurate feature representations, thereby achieving the best performance.

5. Conclusion

In this paper, we propose Flexi-FSCIL, a more flexible semi-supervised framework built upon the prevailing feature-freezing paradigm to address the Few-Shot Class-Incremental Learning (FSCIL) problem. Our method integrates Adaptive Gated Residual Fusion (AGRF) for balanced knowledge retention and adaptive learning, Attention-Guided Dynamic Hybrid Distillation (ADHD) for effective feature alignment, and Prototype Offset Equilibrium (POE) for selective data utilization in stable semi-supervised learning. Experimental results on three benchmark datasets demonstrate that not only achieves state-of-the-art performance but also exhibits strong compatibility with existing feature-freezing models.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant 92267205, in part by the Natural Science Foundation of Hunan Province under Grant 2025JJ60423 and 2025JJ10007, and in part by the Graduate Innovation Program of Central South University (2025ZZTS0323), and in part by the Natural Science Foundation of Shanghai under Grant No. 25ZR1402268,

References

- [1] Noor Ahmed, Anna Kukleva, and Bernt Schiele. Orco: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28762–28771, 2024. 2
- [2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018. 2
- [3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3
- [4] Yawen Cui, Wuti Xiong, Mohammad Tavakolian, and Li Liu. Semi-supervised few-shot class-incremental learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1239–1243. IEEE, 2021. 2, 5, 6
- [5] 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, 2022. 2, 5, 6
- [6] 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*, 2023. 2, 5, 6
- [7] 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*, pages 248–255. Ieee, 2009. 5
- [8] Songlin Dong, Xiaopeng Hong, Xiaoyu Tao, Xinyuan Chang, Xing Wei, and Yihong Gong. Few-shot class-incremental learning via relation knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1255–1263, 2021. 4
- [9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1
- [12] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5070–5079, 2019. 3
- [13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Toronto, ON, Canada, 2009. 5
- [14] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 3
- [15] Jit Yan Lim, Kian Ming Lim, Chin Poo Lee, and Yong Xuan Tan. Ssl-protonet: Self-supervised learning prototypical networks for few-shot learning. *Expert Systems with Applications*, 238:122173, 2024. 2
- [16] Bin Liu, Zhirong Wu, Han Hu, and Stephen Lin. Deep metric transfer for label propagation with limited annotated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3
- [17] Parinita Nema and Vinod K Kurmi. Strategic base representation learning via feature augmentations for few-shot class incremental learning. *arXiv preprint arXiv:2501.09361*, 2025. 5, 6, 7
- [18] Junghun Oh, Sungyong Baik, and Kyoung Mu Lee. Closer: Towards better representation learning for few-shot class-incremental learning. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024. 5, 6
- [19] Mayur Patidar, Riya Sawhney, Avinash Singh, Biswajit Chatterjee, Indrajit Bhattacharya, et al. Few-shot transfer learning for knowledge base question answering: Fusing supervised models with in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9147–9165, 2024. 2
- [20] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017. 2
- [21] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 2, 5, 6
- [22] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 5
- [23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1
- [24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5
- [25] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with

- memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 2
- [26] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in neural information processing systems*, 34: 6747–6761, 2021. 3
- [27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2
- [28] 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. *Advances in neural information processing systems*, 33:596–608, 2020. 3
- [29] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24183–24192, 2023. 2, 5, 6, 7
- [30] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12183–12192, 2020. 5
- [31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7
- [32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 2
- [33] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 5
- [34] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9092–9101, 2021. 3
- [35] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022. 2
- [36] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021. 2
- [37] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023. 2, 5, 6
- [38] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017. 2
- [39] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3
- [40] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2021. 5, 6
- [41] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13208–13217, 2020. 2
- [42] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11838–11847, 2023. 4
- [43] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9046–9056, 2022. 2, 5, 6, 7
- [44] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6801–6810, 2021. 5, 6
- [45] Songhao Zhu and Kai Zhang. Few-shot object detection via data augmentation and distribution calibration. *Machine Vision and Applications*, 35(1):11, 2024. 2