On the Approximation Risk of Few-Shot Class-Incremental Learning

Xuan Wang[1,2], Zhong Ji[1,2]*, Xiyao Liu[3], Yanwei Pang[1,2], and Jungong Han[4]

[1] Tianjin University [2] Shanghai Artificial Intelligence Laboratory
[3] State Key Laboratory of Robotics, SIA, CAS [4] University of Sheffield

Abstract. Few-Shot Class-Incremental Learning (FSCIL) aims to learn new concepts with few training samples while preserving previously acquired knowledge. Although promising performance has been achieved, there remains an underexplored aspect regarding the basic statistical principles underlying FSCIL. Therefore, we thoroughly explore the approximation risk of FSCIL, encompassing both transfer and consistency risks. By tightening the upper bounds of these risks, we derive practical guidelines for designing and training FSCIL models. These guidelines include (1) expanding training datasets for base classes, (2) preventing excessive focus on specific features, (3) optimizing classification margin discrepancy, and (4) ensuring unbiased classification across both base and novel classes. Leveraging these insights, we conduct comprehensive experiments to validate our principles, achieving state-of-the-art performance on three FSCIL benchmark datasets. Code is available at `https://github.com/xwangrs/Approximation_FSCIL-ECCV2024.git`.

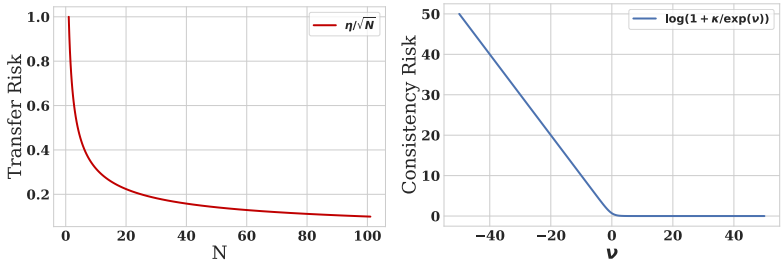
Keywords: Few-Shot Class-Incremental Learning · Few-Shot Learning · Class Incremental Learning · Approximation Risk

1 Introduction

Few-Shot Class-Incremental Learning (FSCIL) [4, 23, 34], an emergent learning paradigm, aims to learn novel classes with limited training data, while preserving the knowledge acquired from base classes. This paradigm faces two primary issues: (1) the catastrophic forgetting of established base class knowledge when integrating new information, and (2) the overfitting to few-shot novel classes.

To effectively address these issues, researchers have proposed various approaches, broadly categorized into three groups [34]: data-based [13, 17, 38], structure-based [23, 32, 36], and optimization-based [2, 4, 35] approaches. Despite their proven effectiveness, there still remains an underexplored aspect regarding the basic statistical principles underlying FSCIL. Central to this gap is the question: **What are the key factors in solving FSCIL problems?** To answer this question, our research delves into the basic statistical principles of FSCIL, aiming to guide the design and training of FSCIL models.

* Corresponding author: Zhong Ji, Email: jizhong@tju.edu.cn



(a) Transfer Risk **(b)** Consistency Risk

Fig. 1: The upper bounds of the transfer and consistency risks. In Fig. 1 (a), the horizontal axis represents the number of training samples for base classes, whereas the vertical axis quantifies the magnitude of the transfer risk. Similarly, in Fig. 1 (b), the horizontal axis reflects the classification margin discrepancy, while the vertical axis quantifies the magnitude of the consistency risk. The parameters $\eta > 0$ and $\kappa > 0$ serve as dynamic scaling factors, regulating the transfer and consistency risks.

We begin with fundamental assumptions: *Homogeneity*, *Regularity conditions*, and *Realizability*. These assumptions, grounded in practical scenarios, form the foundation of our theory.

Based on these assumptions, we formulate the gap between empirically obtained and theoretically optimal FSCIL models as an optimization problem, whose objective function is defined as the **Approximation Risk** of FSCIL. This risk includes two components [1]: *transfer risk* and *consistency risk*. The transfer risk [24] evaluates the efficacy of the shared representations in knowledge transfer from base to novel classes. A low transfer risk indicates effective utilization of base classes knowledge to compensate for the scarcity of novel class samples, thereby preventing overfitting in these novel classes. Meanwhile, the consistency risk measures performance drop when incorporating novel classes, with a low risk signifying robust and stable performance throughout class-incremental tasks, thereby preventing the catastrophic forgetting of base classes.

Leveraging statistical learning theory [6, 14, 21, 26], we explore the upper bounds of transfer and consistency risks, then identifying their determinants[2]:

$$\underbrace{\eta \cdot \frac{1}{\sqrt{N}}}_{\text{Transfer Risk}} + \underbrace{\log(1+\frac{\kappa}{\exp(\nu)})}_{\text{Consistency Risk}}, \tag{1}$$

where N denotes the number of training samples for the base classes; ν indicates the classification margin discrepancy, defined as the infimum of classifier logits for correct classifications minus the supremum of those for incorrect classifications; the parameters η and κ are the dynamic scaling factors for transfer and consistency risks, respectively. We will further discuss and elaborate on methods for minimizing these risks by tightening their upper bounds.

[1] Proofs in Section 4.
[2] Proofs in Section 5.

In Fig. 1 (a), it is evident that increasing the number of training samples for base classes leads to a significant reduction in the transfer risk. Theoretically, this risk could potentially be minimized to a negligible level with a sufficiently large and diverse dataset. However, in practical scenarios, gathering sufficient data for base classes remains a challenge, thereby prompting reliance on foundation models [19,31]. Crucially, for the *Homogeneity* assumption to hold, the marginal distribution of the pre-trained data should align with that of the base class data. Failing to meet this condition could lead to an increase in the transfer risk.

In exploring how N affects the transfer risk, it becomes imperative to fully comprehend the role of the dynamic scaling factor η. From a theoretical perspective, η represents a multivariate polynomial that is related to the maximum activation values across each layer of the neural networks. This could be interpreted as the model's selective attention to specific features. An increase in η values indicates a marked emphasis on certain features, often at the expense of neglecting others, thereby increasing the transfer risk. This phenomenon is known as the "supervision collapse" [3, 9]. To mitigate this issue, we propose to maintain a balanced distribution of activation values across the network layers to prevent extreme impulse values. Furthermore, our theoretical analysis indicates that Vision Transformers (ViTs) [5], owing to their self-attention mechanism [25], exhibit a lower transfer risk compared to Convolutional Neural Networks (CNNs) [8]. Thus, in addressing the FSCIL problem, our preference is towards the deployment of ViT models.

Evidenced in Fig. 1 (b), a clear inverse relationship exists between the classification margin discrepancy and the consistency risk, indicating that enhancing the inter-class discrimination of representations can theoretically diminish the consistency risk to minimal levels. This requires models to maximize the classification margin discrepancy on base classes, ensuring comprehensive training of these models. Moreover, our theory also indicates that the scaling factor κ reaches its lower bound when the model exhibits unbiased classification across both base and novel classes. Achieving this balance involves training classifiers for each task independently before combining them into a unified classifier.

We summarize our contributions as follows:

- Theoretical Framework: We develop a new optimization theory that provides a novel perspective for understanding and addressing challenges in FSCIL.
- Risk Analysis: We decompose the approximation risk of FSCIL into transfer and consistency risks, defining their upper boundaries and and key determinants, thereby offering guidance for model design and training.
- Practical Strategies: We propose practical guidelines to address the challenges in FSCIL. These include expanding training datasets for base classes, preventing excessive focus on specific features, optimizing classification margin discrepancy, and ensuring unbiased classification across both base and novel classes.

Informed by these insights, we achieve state-of-the-art performance on three benchmark datasets, demonstrating the effectiveness of our theory.

2 Related Work

Recently, FSCIL [23,28] has emerged as a paradigm to incrementally learn novel classes with a limited number of samples, while preserving previously learned knowledge. According to the survey [34], existing approaches can be broadly categorized into three groups: data-based approaches [13,17,38], structure-based approaches [23,32,36], and optimization-based approaches [2,4,35].

The data-based approaches address FSCIL from the perspective of training data, which involves data replay-based methods [13] and pseudo scenarios-based methods [17,38]. For example, Kukleva *et al.* [13] propose a three-stage framework where the base and the novel class samples are efficiently replayed to calibrate the classification bias. Zhou *et al.* [38] propose ForwArd Compatible Training (FACT) for FSCIL, which preassigns multiple virtual prototypes in the embedding space by creatively mixuping the known class samples, and Peng *et al.* [17] have similar ideas about utilizing the Mixup [33] to expand the datasets.

The structure-based approaches address FSCIL from the perspective of network architecture design, which involves dynamic structure-based methods [23, 32] and attention-based methods [36]. For example, Zhang *et al.* [32] propose Continually Evolved Classifier (CEC) that employs a graph model to propagate context information between classifiers for adaptation. Tao *et al.* [23] propose to use a neural "gas" network to learn and preserve the topology of the feature manifold formed by the base and the novel classes in FSCIL. Zhao *et al.* [36] propose an attention-based aggregation module that selectively merges predictions from the base branch and the novel branch.

The optimization-based approaches address FSCIL from the perspective of algorithm optimization, which involves representation learning-based methods [35] and knowledge distillation methods [2,4]. Zhao *et al.* [35] concentrate on the dilemma between the slow forgetting of old knowledge and the fast adaptation to novel knowledge. Then a multi-grained "slow vs fast" learning strategy is proposed to cope with this dilemma from both intra-space and the inter-space. Dong *et al.* [4] propose a exemplar relation distillation incremental learning framework to balance the tasks of old-knowledge preserving and new-knowledge adaptation. Cui *et al.* [2] propose to distill reliable knowledge from the reference model, then implement an uncertainty-aware distillation module that combines uncertainty-guided knowledge refinement with adaptive distillation.

Beyond the above approaches, we explore the basic statistical principles underlying FSCIL, then conclude with practical guidelines for the design and training of FSCIL models.

3 Preliminary

Problem Definition. FSCIL involves a base task followed by a sequence of incremental tasks, each identified by a unique task identifier t. Each task is constituted by a distinct training set, $\mathcal{X}^t_{train}$, and a testing set, $\mathcal{X}^t_{test}$, which are associated with their respective distinct label spaces $\mathcal{Y}^t$, such that $\forall i, j$ with

$i \neq j$, $\mathcal{Y}^i \cap \mathcal{Y}^j = \emptyset$. For a particular task t, training is confined to $\mathcal{X}^t_{train}$, while testing is performed across all encountered tasks. Notably, $\mathcal{X}^0_{train}$ encompasses a substantial number of samples for the base task, while $\mathcal{X}^{t>0}_{train}$ are limited to M samples per task.

To tackle FSCIL, popular work [17,30–32] assumes a shared feature representation, enabling knowledge transfer from the base to incremental tasks. Building upon this premise, we introduce the following:

Assumption 1 (Homogeneity) *In FSCIL, the data pairs $\{(\mathbf{x}_{t,i}, y_{t,i})\}$ for each task t, are i.i.d. drawn from a distribution $\mathbb{P}_t$ over $\mathcal{X}^t \times \mathcal{Y}^t$:*

$$\mathbb{P}_t(\mathbf{x}, y) = \mathbb{P}_{\mathbf{x}}(\mathbf{x}) \cdot \mathbb{P}_{y|\mathbf{x}}(y|f_t^\star \circ \mathbf{h}^\star(\mathbf{x})), \tag{2}$$

where $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^r$ denotes a shared feature representation, and $f_t : \mathbb{R}^r \to \mathbb{R}^c$ represents a task-specific mapping. Crucially, this assumption entails that all tasks are subject to an identical marginal distribution $\mathbb{P}_{\mathbf{x}}$, indicative of a unified dataset shared across the tasks.

4 Approximation Risk of FSCIL Models

As explicated in existing work [23,32], the training procedure for FSCIL models is roughly partitioned into two distinct phases: the base and incremental phases.

Base Phase. During the base phase, the primary objective is the learning of an optimal shared representation $\mathbf{h}$ in conjunction with the base classifier f_0. This is accomplished by minimizing the empirical risk $\hat{R}_{\text{base}}$:

$$\hat{R}_{\text{base}}(f_0, \mathbf{h}) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_0 \circ \mathbf{h}(\mathbf{x}_{0,i}), y_{0,i}), \tag{3}$$

where $\ell(\cdot, \cdot)$ symbolizes the loss function; $\mathbf{x}_{0,i} \in \mathcal{X}^0_{\text{train}}$ and $y_{0,i} \in \mathcal{Y}^0$. The composition operator $\circ$ denotes the consecutive application of functions such that $f_0 \circ \mathbf{h}(\mathbf{x}) = f_0(\mathbf{h}(\mathbf{x}))$. If Assumption 1 holds and the dataset size N is sufficiently large ($N \gg 1$), the estimator $(\hat{f}_0, \hat{\mathbf{h}})$, derived from the minimization of $\hat{R}_{\text{base}}$, serves as an effective proxy for the optimal function pair $(f_0^\star, h^\star)$. This base phase is crucial in forming a robust baseline for the learning agent, enabling a seamless transition and effective integration with the novel information presented in subsequent incremental learning tasks.

Incremental Phase. The incremental training process continues with a sequence of incremental tasks, where the learning agent learns the task-specific classifier f_t by minimizing empirical risk $\hat{R}_{\text{novel}}$ over M new training samples [3]:

$$\hat{R}_{\text{novel}}([f_0, f_t], \mathbf{h}) = \frac{1}{M} \sum_{m=1}^{M} \ell([f_0, f_t] \circ \mathbf{h}(\mathbf{x}_{t,m}), y_{t,m}), \tag{4}$$

[3] For brevity in our theoretical exposition, we confine this discussion to a single incremental task; nevertheless, the principles delineated here are also applicable to sequential tasks.

where $[f_0, f_t]$ signifies the concatenation of the base and incremental classifiers; $\mathbf{x}_{t,m} \in \mathcal{X}^t_{\text{train}}$ and $y_{t,m} \in \mathcal{Y}^t$. The challenge herein lies in the fact that for $M \ll N$, the estimator $([\hat{f}_0, \hat{f}_t], \hat{\mathbf{h}})$, obtained from minimizing $\mathbf{h} \in \mathcal{H}$ fails to estimate their underlying counterparts $([f_0^\star, f_t^\star], h^\star)$. To address this limitation, we assess the efficacy of the estimator $([\hat{f}_0, \hat{f}_t], \hat{\mathbf{h}})$ by its excess risk on the incremental task t, defined as the approximation risk of FSCIL:

$$\Delta_{\text{AR}} = R_{\text{novel}}([\hat{f}_0, \hat{f}_t], \hat{\mathbf{h}}) - R_{\text{novel}}([f_0^\star, f_t^\star], \mathbf{h}^\star). \tag{5}$$

Approximation Risk. To clarify the approximation risk, we deconstruct $\Delta_{\text{AR}} = \Delta_{\text{transfer}} + \Delta R_{\text{novel}}$[4]. These components are defined as:

$$\Delta_{\text{transfer}} = R_{\text{novel}}(\hat{f}_t, \hat{\mathbf{h}}) - R_{\text{novel}}(f_t^\star, \mathbf{h}^\star), \tag{6}$$

$$\Delta R_{\text{novel}} = R_{\text{novel}}([\hat{f}_0, \hat{f}_t], \hat{\mathbf{h}}) - R_{\text{novel}}(\hat{f}_t, \hat{\mathbf{h}}). \tag{7}$$

Within this framework, Δ_{AR} is deconstructed into two principal components: Δ_{transfer} and ΔR_{novel}. The former signifies the transfer risk, assessing the expected prediction risk when transferring knowledge to novel tasks. The latter encapsulates the adaptation risk, which particularly measures the divergence between the composite classifier $[f_0, f_t]$ and the task-specific classifier f_t.

Beyond Δ_{AR}, FSCIL additionally necessitates the consistent performance across incremental tasks. This is formally represented by a constraint encapsulating the adaptation risk pertaining to the base classifier:

$$\Delta R_{\text{base}} = R_{\text{base}}([\hat{f}_0, \hat{f}_t], \hat{\mathbf{h}}) - R_{\text{base}}(\hat{f}_0, \hat{\mathbf{h}}). \tag{8}$$

In synthesis, the final version of the approximation risk is represented by Eq. (5), incorporating the constraint given by Eq. (8), and is achieved by applying the penalty method:

$$\min_{f_0, f_t, \mathbf{h}} \underbrace{\Delta_{\text{transfer}}}_{\text{Transfer Risk}} + \underbrace{\Delta R_{\text{novel}} + \gamma \Delta R_{\text{base}}}_{\text{Consistency Risk}}, \tag{9}$$

where γ represents the penalty coefficient.

5 Theoretical Results And Applications

In this section, the basic statistical principles underlying FSCIL are presented. Initially, we make the following standard, mild regularity assumptions on the loss function $\ell(\cdot, \cdot)$, the function class of tasks $\mathcal{F}$, and the function class of shared representations $\mathcal{H}$.

Assumption 2 (Regularity conditions) *The following conditions hold:*

- *The loss function $\ell(\cdot, \cdot)$ is B-bounded, and $\ell(\cdot, y)$ is Γ-Lipschitz for all $y \in \mathcal{Y}$.*

[4] Proof In Appendix 9.4

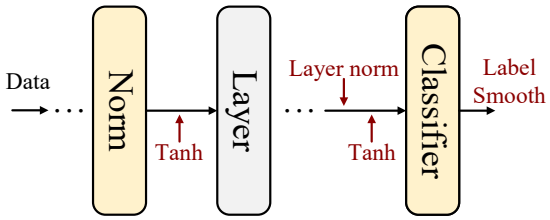


Fig. 2: Architecture-Agnostic Modifications to FSCIL Models. Specific adjustments are indicated in red.

- *The function f is $\Gamma(\mathcal{F})$-Lipschitz regarding the ℓ_2 distance for any $f \in \mathcal{F}$.*
- *The composed function $f \circ \mathbf{h}$ is bounded:* $\sup_{\mathbf{x}\in\mathcal{X}} |f \circ \mathbf{h}(\mathbf{x})| \leq D_{\mathcal{X}}$*, for any* $f \in \mathcal{F}, \mathbf{h} \in \mathcal{H}$.

Furthermore, the realizability assumption is adopted, positing that the true task functions and the true representation are contained in $\mathcal{F}$ and $\mathcal{H}$, respectively.

Assumption 3 (Realizability) *The true representation $\mathbf{h}^\star$ is encompassed by $\mathcal{H}$. Concurrently, for all tasks considered—both base and incremental—the true task-specific functions $f_t^\star$ are also contained in $\mathcal{F}$.*

Based on Regularity and Realizability assumptions, we analyze transfer learning risk Δ_{transfer} and consistency risk $\Delta R_{\text{novel}} + \gamma \Delta R_{\text{base}}$ in Eq. (9) separately.

5.1 Transfer Risk

Theorem 1. *(Proof in Appendix 9.2) If Assumptions 2 and 3 hold, and the base task is $(\rho, 0)$ diverse, with probability $1 - 2\delta$, the transfer risk $\Delta_{transfer}$ in Eq. (9) is bounded by:*

$$\tilde{O}\Big(\frac{\Gamma}{\rho} \cdot \Gamma(\mathcal{F}) \cdot 4D \prod_{l=1}^{L} Z(l) \sqrt{\frac{\log 2J}{N}} + \frac{\Gamma D_{\mathcal{X}^0}}{\rho N^2} + B\Big[\frac{1}{\rho} \cdot \sqrt{\frac{\log(2/\delta)}{N}} + \sqrt{\frac{\log(2/\delta)}{M}}\Big]\Big),$$

where N and M quantify the number of training samples of the base and incremental task t respectively; $\|\mathbf{x}\| \leq D$; L is the depth of the neural network; $Z(l) = \frac{\|h^{(l)}\|_\infty}{\|h^{(l-1)}\|_\infty}$, where $h^{(l)}$ denotes the feature representations of the l-th layer.

Application. To bridge the basic statistical principles and the practical guidelines, we undertake the subsequent analysis:

(1) From Theorem 1, it becomes apparent that the quantity N, representing the number of training samples for the base task, primarily influences the upper bound of the transfer risk. As N increases, the upper bound tightens at a rate proportional to $O(\frac{1}{\sqrt{N}})$. This relation shows that **expanding training datasets for base classes** is the most straightforward and potent strategy for addressing the FSCIL challenge, especially when the initial dataset is insufficiently large.

In scenarios where expanding the data pool is not feasible, the adoption of a foundation model [19,27] becomes a strategic alternative, which is substantiated by our analysis of the transfer risk.

(2) Theoretically, the upper bound of transfer risk is proportional to $\prod_{l=1}^{L} Z(l)$, where $Z(l)$ is the ratio of the infinity norm of features at layer l to that at layer $l-1$. While reducing $\prod_{l=1}^{L} Z(l)$ could lower transfer risk, it might impair the feature extraction, so we do not adopt this modification. In original ResNets and ViTs, each layer typically includes Layer or Batch Normalization, standardizing feature means to 0 and variances to 1. However, this normalization does not limit the infinity norm of features, potentially causing excessively large values in some dimensions and increasing $Z(l)$, then raising transfer risk. To mitigate this, we constrain each layer's features to $[-1, 1]$, setting $Z(l) = 1$. Since normalization centers most feature values around zero, we introduce the *tanh* function to cap them. The *tanh* function bounds values within $[-1, 1]$ and has a slope of 1 at zero, effectively constraining extreme values without significantly impacting original models. Similarly, we also use *norm* + *tanh* to constrain the representation $\mathbf{h}$.

Besides, $D_{\mathcal{X}^0}$ represents the highest probability value of the predicted outcome, exhibiting a positive correlation with the upper bound. The essential strategy to reduce this bound involves calibrating the model to ensure its predictions are tempered, thereby preventing overconfidence. In conclusion, the analysis of both $\prod_{l=1}^{L} Z(l)$ and $D_{\mathcal{X}^0}$ shows that **preventing excessive focus on specific features** is an effective strategy for addressing FSCIL challenge. As shown in Fig. 2, we propose three architecture-agnostic modifications:

- Incorporation of the *tanh* activation function subsequent to each normalization layer in the original models;
- Integration of the layer normalization and *tanh* activation function prior to the generation of shared representations;
- Application of the label smooth in the training process.

Remark. The term $Z(l) = \frac{\|h^{(l)}\|_\infty}{\|h^{(l-1)}\|_\infty}$ quantifies the ratio of maximal activation value in the l-th layer compared to that in the $(l-1)$-th layer. It serves as an indicator of the proportional activation dynamics spanning all dimensions between consecutive layers. This metric emphasizes the need for a balanced distribution of activation values throughout the network layers to inhibit the incidence of impulse activation. Such regulation ensures that the model maintains an equitable emphasis on all features, thereby supporting the underlying expectation of $D_{\mathcal{X}^0}$ for the model to prevent overconfidence in its predictions.

Corollary 1 (Proof in Appendix 9.3). *Under conditions from Theorem 1, Vision Transformers (ViTs) exhibit lower transfer risks than Convolutional Neural Networks (CNNs).*

Remark. This finding substantiates the efficacy of ViTs in leveraging shared representations for novel tasks, underscoring their robustness in transfer learning scenarios. Furthermore, the reduced transfer risk with ViTs as compared to CNNs suggests a pivotal architectural advantage that may inform future developments in neural network design.

5.2 Consistency Risk

Theorem 2 (Proof in Appendix 9.5). *If Assumptions 2 and 3 hold, and given K_0 and K_t classes in the base and incremental tasks respectively, the consistency risk $\Delta R_{novel} + \gamma \Delta R_{base}$ in Eq. (9) is upper-bounded by:*

$$\log\left((1+\frac{K_0}{K_t}v_t)(1+\frac{K_t}{K_0}v_0)^{\gamma}\right), \tag{10}$$

where v_t and v_0 involve supremum and infimum of logits:

$$v_t = \exp\left(\sup(\phi_{t\to 0}) - \inf(\phi_{t\to t})\right), \tag{11}$$

$$v_0 = \exp\left(\sup(\phi_{0\to t}) - \inf(\phi_{0\to 0})\right). \tag{12}$$

The classifier logits $\phi_{t\to 0}$, $\phi_{t\to t}$, $\phi_{0\to 0}$, and $\phi_{0\to t}$ respectively quantify how samples from task t and the task 0 (base task) are classified, either correctly or in-correctly.

Application. Building on the above analysis about the transfer risk, we proceed with the following detailed analysis about the consistency risk:

(3) From Theorem 2, we observe that the consistency risk is inherently tied to the **classification margin discrepancy** of the model, which is quantified by the difference between the infimum of the classifier logits for correct classification and the supremum for incorrect classification. It infers that the consistency risk approaches nullity as the predictive accuracy of the model approaches perfection.

(4) The upper bound of the consistency risk represents a balance between v_t and v_0. Introducing a bias in categorization leads to inverse variations in v_t and v_0, indicating that **ensuring unbiased classification across both base and novel classes** can completely nullify the consistency risk. This implies strategically designed biases might reach the minimal risk bound but are not optimally efficient. Key lies in improving classification margin discrepancy while maintaining a balanced classification bias.

6 Experiments

6.1 Datasets and Implementation Details

Datasets. Our study employs three diverse datasets: *mini*-ImageNet [20], CIFAR100 [12], and CUB200 [29]. The *mini*-ImageNet, a subset of ImageNet [20], includes 600 images from 100 classes, split into 60 base and 40 novel classes. These novel classes are divided into 8 incremental tasks, each a 5-way 5-shot task. CIFAR100 comprises 100 classes with 600 images each, partitioned into 60 base and 40 novel classes, the latter organized into 8 incremental 5-way 5-shot tasks. CUB200, containing 11,788 images of 200 bird species, is split into 100 base and 100 novel classes, with the latter arranged into 10 incremental 10-way 5-shot tasks. In this work, when the model is trained from scratch, the image sizes for *mini*-ImageNet, CIFAR100, and CUB200 are set at 84×84, 32×32,

Table 1: Classification Accuracy (%) on CUB200 in the 10-Way 5-Shot FSCIL Setting. The notation ViT-B/16 specifically denotes the ViT-Base model [5] with a designated patch size of 16. The symbols ⋄ and † signify models derived from the CLIP [19] and those pre-trained on the ImageNet-1K dataset respectively. The symbol ↓ is employed to indicate that lower values are more desirable.

Method	Backbone	Accuracy in each session(%)											DR↓(%)
		0	1	2	3	4	5	6	7	8	9	10	
CEC† [32]	ResNet18	75.85	71.94	68.50	63.5	62.43	58.27	57.73	55.81	54.83	53.52	52.28	31.07
MCNet† [10]	ResNet18	77.57	73.96	70.47	65.81	66.16	63.81	62.09	61.82	60.41	60.09	59.08	23.84
FACT† [38]	ResNet18	75.90	73.23	70.84	66.13	65.56	62.15	61.74	59.83	58.41	57.89	56.94	24.98
ALICE† [17]	ResNet18	77.4	72.7	70.6	67.2	65.9	63.4	62.9	61.9	60.5	60.6	60.1	22.35
CABD† [37]	ResNet18	79.12	75.63	73.21	69.93	68.32	66.30	65.15	64.96	64.20	64.03	63.81	19.35
NC-FSCIL† [30]	ResNet18	80.45	75.98	72.30	70.28	68.17	65.16	64.43	63.25	60.66	60.01	59.44	26.12
SAVC† [22]	ResNet18	81.85	77.92	74.95	70.21	69.96	67.02	66.16	65.30	63.84	63.15	62.50	23.64
LDC† [15]	ResNet18	77.89	76.93	74.64	70.06	68.88	67.15	64.83	64.16	63.03	62.39	61.58	20.94
CoCoOp⋄ [39]	ViT-B/16	80.3	77.1	75.8	73.9	72.4	68.2	65.2	63.9	61.1	60.3	57.2	28.76
Prompt⋄ [31]	ViT-B/16	84.3	83.2	80.9	79.2	77.7	72.4	72.1	69.9	68.8	67.4	66.8	20.76
ours†	ResNet18	79.87	77.36	74.92	70.62	70.64	68.55	68.14	67.93	66.73	66.67	65.69	17.75
ours⋄	ViT-B/16	86.46	84.21	82.67	79.79	79.52	77.32	76.71	77.16	75.76	75.56	75.22	13.00
ours†	ViT-B/16	**86.69**	**85.38**	**84.34**	**82.43**	**83.08**	**81.07**	**81.38**	**81.42**	**81.21**	**81.11**	**81.30**	**6.22**

and 224×224 respectively. In scenarios utilizing a foundation model, the images from all three datasets are resized to 224×224.

Implementation Details. Following previous work [23, 34], and guided by our theory, we select ViTs as our primary models across all datasets. Specifically, we utilize two ViT-B/16 variants: one from CLIP [19] and another pre-trained on ImageNet-1K. During the base phase, we train the ViT models on the base classes for 50 epochs using the AdamW optimizer. The learning rate is initially set to 0, then linearly rises to 0.00001 during the first 20 epochs, followed by a linear decrease back to 0 for the rest of the epochs. For a fair comparison, ResNet18 serves as an auxiliary model across datasets, trained on base classes for 500 epochs using AdamW. The learning rate for ResNet18 follows a similar pattern: beginning at 0, increasing linearly to 0.1 within the first 50 epochs, and subsequently diminishing back to 0. In the incremental phase, we freeze the backbone and update only the task-specific classifiers.

Evaluation Metrics. Following precedents [16, 23], our evaluation employs Top-1 classification accuracy across cumulative testing sets after each learning session. While traditional models often employ the Performance Drop (PD) metric, defined as $\mathrm{PD} = (\mathcal{A}_0 - \mathcal{A}_N)$, where $\mathcal{A}_0$ and $\mathcal{A}_N$ represent the accuracies at the initial and final sessions, we propose a refined metric. Recognizing that model forgetting encompasses both the volume of initially acquired knowledge and the extent of subsequent knowledge loss, we introduce the Performance Drop Rate (DR), calculated as $\mathrm{DR} = (\mathcal{A}_0 - \mathcal{A}_N)/\mathcal{A}_0$. This metric more accurately captures the effect of knowledge loss, factoring in the initial knowledge base.

6.2 Comparison with State-of-the-Art Methods

We benchmark our model against state-of-the-art performances on three datasets: CUB200, CIFAR100, and *mini*-ImageNet, with results in Fig. 3 and detailed CUB200 data in Table 1 (see supplementary materials for other datasets).

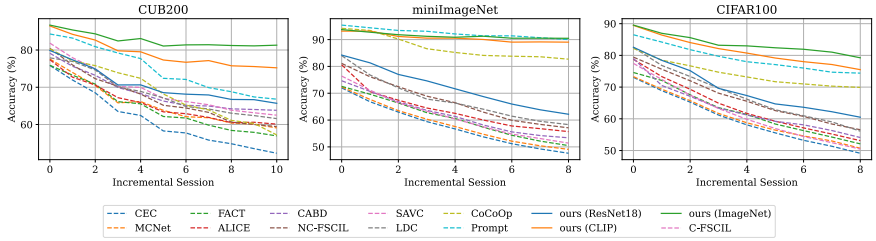


Fig. 3: Comparison with State-of-the-Art on Three Datasets: CUB200, *mini*-ImageNet, and CIFAR100. For detailed performance metrics, please see Table 1 and refer to our supplementary material.

Table 1 shows that our ViT model (CLIP) achieves an improvement of 8.42% in the final session compared to the sub-optimal method, and the DR metric decreases by 7.76%. When ImageNet-1K is used for pre-training, our ViT model records a boost of 17.49% in the last session compared to the second best method, with a DR reduction of 13.13%. As for ResNet18, our ResNet18 model exhibits a 1.88% improvement in the final session over the sub-optimal ResNet18 model, with a 1.6% reduction in the DR metric. These results indicate a decrease in catastrophic forgetting and overfitting, attributed to the application of our guiding principles for reducing the approximation risk.

From Fig. 3, similar trends are observed on the *mini*-ImageNet and CIFAR100, though the improvements of ViTs are less pronounced. This observation can be attributed to *mini*-ImageNet being a subset of ImageNet, which violates the standard FSCIL setting, and to the unique image characteristics of CIFAR100, such as smaller image size and lack of background, which differ from the typical natural image distributions used in pre-training.

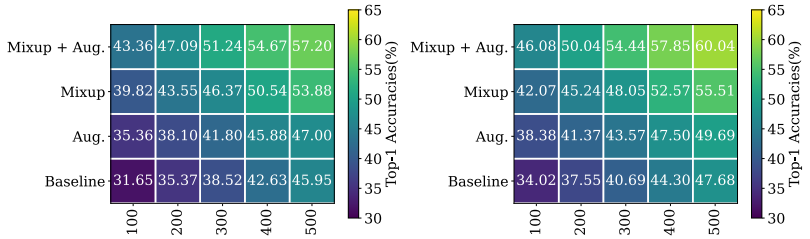
Remarkably, even though CLIP has a broader pre-training dataset compared to ImageNet-1K, performance with CLIP remains consistently lower. This phenomenon might be explained by the significant difference between the marginal distributions of the three benchmark datasets and the data used for pre-training of CLIP. Conversely, ImageNet-1K aligns more closely with these distributions, better fulfilling the Homogeneity assumption in our study.

6.3 Ablation Study

In this section, we assess the impact of four key components of our ViT model on the CUB200 dataset (see supplementary materials for ResNet18). Detailed in Table 2, these components show varied performance effects. The FT strategy exhibits the lowest Top-1 accuracy (38.01% in the final session) and highest DR metric (40.99%), indicating a pronounced bias towards novel classes and substantial consistency risk. Conversely, FR strategy results in improved Top-1 accuracy (52.41%) and lower DR metric (36.63%), yet struggles with catastrophic forgetting. The DS strategy, separating training of base and novel classifiers, is the most effective in reducing bias and consistency risk.

Table 2: Ablation Study on CUB200 in the 10-Way 5-Shot FSCIL Setting. With ViT-B/16 as the backbone, our study encompasses: "Stabilize", enhacing output stability via *tanh* function after normalization; "Constrain", limiting representation values using layer normalization and *tanh* function; "LS" (Label Smoothing) to reduce model overconfidence; and "BC" (Base Classifier), employing Fine-Tuning (FT), Freezing (FR), and Discarding (DS) in the incremental phase.

Stabilize	Constrain	LS	BC	Accuracy in each session(%)											DR↓(%)
				0	1	2	3	4	5	6	7	8	9	10	
			FT	64.42	59.32	55.04	50.62	48.91	45.22	43.97	42.13	40.59	39.50	38.01	40.99
			FR	82.71	76.36	71.33	66.95	64.88	61.03	58.94	57.48	55.75	54.12	52.41	36.63
			DS	87.05	83.50	81.00	77.66	76.56	74.12	73.00	73.17	72.32	71.87	70.97	18.47
✓			DS	87.15	84.41	81.65	78.11	77.46	75.18	74.44	73.87	73.17	72.97	72.07	17.30
	✓		DS	86.77	84.02	81.56	78.31	77.63	75.44	74.39	74.48	73.40	73.11	72.29	16.69
		✓	DS	87.49	84.09	82.19	78.56	77.93	75.49	75.53	74.67	73.46	73.42	72.77	16.82
✓	✓		DS	87.51	84.03	81.92	78.43	78.22	75.74	75.91	75.19	73.59	73.83	73.34	16.19
✓		✓	DS	**87.63**	84.13	81.77	78.05	78.00	75.80	75.40	74.79	74.43	74.25	73.71	15.88
	✓	✓	DS	87.46	**84.53**	**82.77**	79.32	78.96	76.62	**76.85**	75.87	74.54	74.47	73.98	15.41
✓	✓	✓	DS	86.46	84.21	82.67	**79.79**	**79.52**	**77.32**	76.71	**77.16**	**75.76**	**75.56**	**75.22**	**13.00**



(a) ResNet18 **(b)** ViT-S

Fig. 4: Impact of Dataset Size in Transfer Risk with ResNet18 and ViT-S/6 Backbones Trained from Scratch on *mini*-ImageNet. In both subfigures, the horizontal axis represents the dataset size per base class, while the vertical axis lists training techniques. These include Data Augmentation [1, 7, 11] and Mixup [18, 33], essential for expanding the training dataset.

Integration of the *tanh* function for layer stabilization improves the DR metric by about 1%, mitigating overfitting. Similarly, constrained feature representations through layer normalization and *tanh* function slightly decrease base class accuracy but enhance the DR metric by 1%, indicating reduced overfitting. Implementing label smoothing technique raises both base classification accuracy and DR metric by 1%, attributed to lowering transfer risk.

Combining these components yields superior performance, highlighting their synergistic effect in FSCIL.

6.4 Dataset Size in Transfer Risk

In this section, we empirically investigate the correlation between dataset size and transfer risk in neural networks, focusing on ResNet18 and ViT-S/6 architectures. Fig. 4 (a) and (b), show a clear correlation between increased dataset size per class and improved final accuracy for both models. This supports the

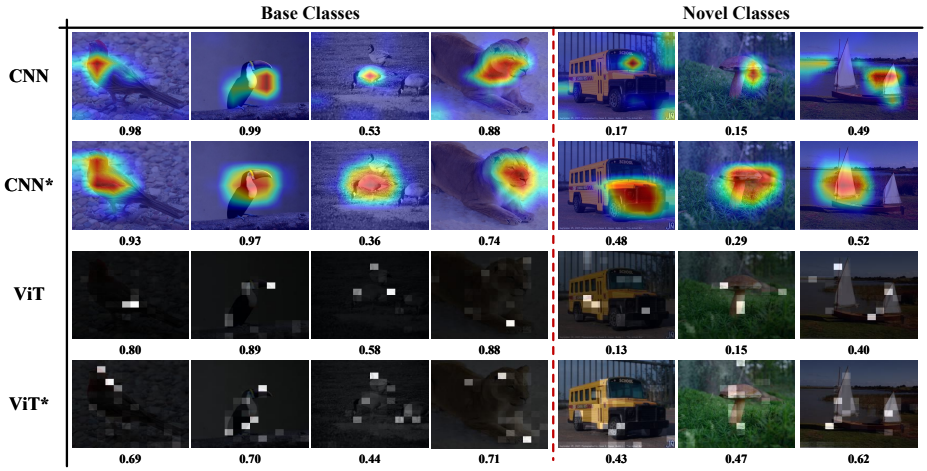


Fig. 5: Visualization of ResNet18 and ViT-S/6 on *mini*-ImageNet. Images are divided by a red dotted line: left for base classes and right for novel classes. Beneath each image, the predicted probability is numerically displayed. The star symbol denotes models in our implementation. In ViT visualizations, image brightness reflects attention level. Darker images signify attention concentrated on specific patches, while brighter ones indicate a more global attention on the whole image.

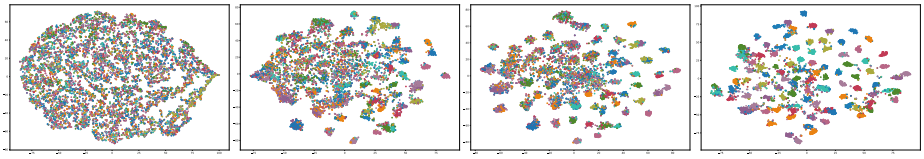
idea that increasing the dataset size (N) effectively tightens the upper bound of transfer risk, thereby enhancing models' capacity for incremental learning.

When direct sample collection is limited, Data Augmentation and Mixup techniques present a viable alternative. These methods create additional samples, enriching the training dataset. This is evident in Figures 4 (a) and (b), where both ResNet18 and ViT-S models show significant performance gains with these techniques. This reinforces our theory: indirect sample generation via Data Augmentation and Mixup can enhance model transferability, mirroring the advantages of direct sample collection.

6.5 Analysis of the Model Overconfidence

In this section, we provide a visual results to illustrate the influence of prediction confidence on transfer risk. The first and third rows of Fig. 5 indicate that models (CNNs or ViTs) naturally focus on small, highly discriminative regions within base class samples, which leads to overconfidence in these predictions. This propensity leads to misclassification during the transfer of models to novel classes, a phenomenon commonly referred to as supervision collapse.

Our approach aims to temper the model's over-reliance on base class predictions. As evidenced in the second and fourth rows of Fig. 5, lowering the probability for base class predictions enables the model to cover broader regions. This expanded scope assists in focusing on distinctive features of novel classes, thus enhancing prediction accuracy for these classes.



(a) -30.84 **(b)** -15.78 **(c)** -10.18 **(d)** -3.5

Fig. 6: The t-SNE visualizations of the logits for test samples on *mini*-ImageNet, under varying degrees of Classification Margin Discrepancy, are depicted with distinct colors representing different classes. The architectures for Fig. (a), (b), and (c) are based on ViT-S/6 models, each trained from scratch. Conversely, Fig. (d) employs a ViT-B/16 backbone, which is pre-trained on the ImageNet dataset.

6.6 Classification Margin Discrepancy

As demonstrated in Fig. 6, we observe a progressive enhancement in the separability of logits for test samples, correlating with an incremental increase in the Classification Margin Discrepancy. This trend not only signifies a reduction in the consistency risk but also marks a significant step towards achieving more reliable and robust classification performance in FSCIL settings. These findings empirically validate our theoretical framework, suggesting that optimizing for classification margin discrepancy directly contributes to mitigating consistency risks and enhancing model adaptability across incremental learning tasks.

7 Conclusion

In conclusion, we have conducted a thorough exploration on the approximation risk of FSCIL. By focusing on mitigating transfer and consistency risks and tightening their upper bounds, we have developed practical guidelines for the design and training of FSCIL models. These guidelines have expanded training datasets for base classes, optimized classification margin discrepancy, prevented excessive focus on specific features, and ensured unbiased classification across both base and novel classes. Our extensive experiments have validated these principles, culminating in state-of-the-art performance on three benchmark datasets.

8 Limitation

In this paper, *Homogeneity* and *Regularity conditions* are foundational assumptions. *Homogeneity*, based on the *i.i.d* assumption, indicates samples in a dataset share similar characteristics like hues and lighting. *Regularity* involves boundedness for model convergence and the Lipschitz condition, crucial for representation learning. Practically, comparative experiments show that even when the pre-training and FSCIL benchmark datasets have different marginal distributions, our methods still achieve better results, albeit with smaller gains. This indicates that our theory generally holds but comes at the cost of some practicality.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160403, by the National Natural Science Foundation of China (NSFC) under Grants 62176178, 62106152, and 62303447, by the Postdoctoral Innovation Talents Support Program under No. BX20230399, and by the China Postdoctoral Science Foundation under No. 2023M743702.

References

1. Balestriero, R., Bottou, L., LeCun, Y.: The effects of regularization and data augmentation are class dependent. In: Advances in Neural Information Processing Systems. vol. 35, pp. 37878–37891 (2022)
2. Cui, Y., Deng, W., Chen, H., Liu, L.: Uncertainty-aware distillation for semi-supervised few-shot class-incremental learning. IEEE Transactions on Neural Networks and Learning Systems (2023), early Access
3. Doersch, C., Gupta, A., Zisserman, A.: Crosstransformers: spatially-aware few-shot transfer. In: Advances in Neural Information Processing Systems. pp. 21981–21993 (2020)
4. Dong, S., Hong, X., Tao, X., Chang, X., Wei, X., Gong, Y.: Few-shot class-incremental learning via relation knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1255–1263 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
6. Golowich, N., Rakhlin, A., Shamir, O.: Size-independent sample complexity of neural networks. In: Conference On Learning Theory. pp. 297–299. PMLR (2018)
7. Hao, X., Zhu, Y., Appalaraju, S., Zhang, A., Zhang, W., Li, B., Li, M.: Mixgen: A new multi-modal data augmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 379–389 (2023)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
9. Hiller, M., Ma, R., Harandi, M., Drummond, T.: Rethinking generalization in few-shot classification. In: Advances in Neural Information Processing Systems. pp. 1–13 (2022)
10. Ji, Z., Hou, Z., Liu, X., Pang, Y., Li, X.: Memorizing complementation network for few-shot class-incremental learning. IEEE Transactions on Image Processing **32**, 937–948 (2023)
11. Ji, Z., Jiao, Z., Wang, Q., Pang, Y., Han, J.: Imbalance mitigation for continual learning via knowledge decoupling and dual enhanced contrastive learning. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2024), early Access
12. Krizhevsky, A., Nair, V., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)

13. Kukleva, A., Kuehne, H., Schiele, B.: Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9020–9029 (2021)
14. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: isoperimetry and processes, vol. 23. Springer Science & Business Media (1991)
15. Liu, B., Yang, B., Xie, L., Wang, R., Tian, Q., Ye, Q.: Learnable distribution calibration for few-shot class-incremental learning. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–8 (2023)
16. Mazumder, P., Singh, P., Rai, P.: Few-shot lifelong learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2337–2345 (2021)
17. Peng, C., Zhao, K., Wang, T., Li, M., Lovell, B.C.: Few-shot class-incremental learning from an open-set perspective. In: European Conference on Computer Vision. pp. 382–397. Springer (2022)
18. Pinto, F., Yang, H., Lim, S.N., Torr, P., Dokania, P.: Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In: Advances in Neural Information Processing Systems. vol. 35, pp. 14608–14622 (2022)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
21. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press (2014)
22. Song, Z., Zhao, Y., Shi, Y., Peng, P., Yuan, L., Tian, Y.: 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. pp. 24183–24192 (June 2023)
23. Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., Gong, Y.: Few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12183–12192 (2020)
24. Tripuraneni, N., Jordan, M., Jin, C.: On the theory of transfer learning: The importance of task diversity. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7852–7862 (2020)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, pp. 1–11 (2017)
26. Wainwright, M.J.: High-dimensional statistics: A non-asymptotic viewpoint, vol. 48. Cambridge university press (2019)
27. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14408–14419 (2023)
28. Wang, X., Ji, Z., Pang, Y., Yu, Y.: A cognition-driven framework for few-shot class-incremental learning. Neurocomputing **600**, 128118 (2024)
29. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)

30. Yang, Y., Yuan, H., Li, X., Lin, Z., Torr, P., Tao, D.: Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In: The Eleventh International Conference on Learning Representations. pp. 1–13 (2023)
31. Yoon, I.U., Choi, T.M., Lee, S.K., Kim, Y.M., Kim, J.H.: Image-object-specific prompt learning for few-shot class-incremental learning. arXiv preprint arXiv:2309.02833 (2023)
32. Zhang, C., Song, N., Lin, G., Zheng, Y., Pan, P., Xu, Y.: Few-shot incremental learning with continually evolved classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12455–12464 (2021)
33. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
34. Zhang, J., Liu, L., Silven, O., Pietikäinen, M., Hu, D.: Few-shot class-incremental learning: A survey. arXiv preprint arXiv:2308.06764 (2023)
35. Zhao, H., Fu, Y., Kang, M., Tian, Q., Wu, F., Li, X.: Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–11 (2021)
36. Zhao, L., Lu, J., Xu, Y., Cheng, Z., Guo, D., Niu, Y., Fang, X.: Few-shot class-incremental learning via class-aware bilateral distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11838–11847 (2023)
37. Zhao, L., Lu, J., Xu, Y., Cheng, Z., Guo, D., Niu, Y., Fang, X.: Few-shot class-incremental learning via class-aware bilateral distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11838–11847 (June 2023)
38. Zhou, D.W., Wang, F.Y., Ye, H.J., Ma, L., Pu, S., Zhan, D.C.: Forward compatible few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9046–9056 (2022)
39. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)