This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Enhancing Few-Shot Class-Incremental Learning via Frozen Feature Augmentation

Shimou Ling*, Shengkai Gan*, Caoxin Wang, Lili Pan†, Hongliang Li
University of Electronic Science and Technology of China
Chengdu, China
lilipan@uestc.edu.cn

Abstract

Few-shot class incremental learning (FSCIL) poses a significant challenge, as it not only requires mitigating catastrophic forgetting but also addressing overfitting. Previous studies have primarily focused on alleviating these issues through representation learning, with the goal of enhancing the separation between base classes and providing adequate representation space for incremental classes. As pre-trained models become increasingly prevalent in continual learning, exploring frozen feature augmentation for FSCIL becomes extremely necessary. This work proposes an ingenious brightness-based frozen feature augmentation method for FSCIL. We validate its effectiveness on both regular and multi-modal continual learning datasets, including CIFAR-100, mini-ImageNet, UESTC-MMEA-CL, and ARIC. Not only that, we analyze the decision attribution stability change afforded by such a feature augmentation, demonstrating that it does not compromise interpretability. Our code is available at https://github.com/learninginvision/FAOrCo-ViT.

1. Introduction

Continual learning represents a pivotal paradigm in artificial intelligence, enabling systems to emulate the human-like capacity for continual knowledge acquisition [31]. Within this sophisticated research domain, few-shot class incremental learning (FSCIL) [28] emerges as a particularly formidable challenge. Diverging fundamentally from conventional continual learning approaches that predominantly rely on extensive training data [4, 13, 17, 24, 26], FSCIL endeavors to synthesize knowledge about emergent classes through strategic integration of minimally available sample instances [12, 30, 35, 36, 39, 40]. This nuanced approach demands sophisticated algorithmic strategies to effectively generalize and expand learning capabilities under data-constrained scenarios.

Existing methods [12, 35, 36, 39] in FSCIL predominantly concentrate on representation learning strategies, with the primary objective of preserving base class representations while facilitating the learning of novel class features. While addressing the critical challenge of overfitting, some literature has predominantly relied on data augmentation techniques [20, 42]; however, feature augmentation, as a potential strategy, remains conspicuously unexplored.

As pre-trained models become more prevalent in continual learning, particularly the Vision Transformer (ViT), gain prominence due to their superior visual recognition performance [32, 41], exploring how to augment the output features of the pre-trained ViT (*i.e.*, frozen features) becomes essential in FSCIL. During few-shot incremental learning sessions, the diversity and the size of the training set can be enhanced via frozen feature augmentation, thereby boosting continual learner's generalization.

To this end, this work proposes a novel brightness-based frozen feature augmentation method for FSCIL. Fig. 1 provides an overview of our method. We integrate this feature augmentation method with one typical FSCIL framework, OrCo [1]. The experimental results show such a feature augmentation improves the few-shot continual learner's accuracy by enhancing the diversity of the features extracted by pre-trained model. Our contributions include:

- Leveraging the architectural foundation of Vision Transformer (ViT), we introduce a novel frozen feature augmentation methodology for FSCIL, demonstrating performance improvements across standard continual learning benchmarks, specifically CIFAR100 and mini-ImageNet.
- Our proposed FSCIL approach is comprehensively validated through extension to two open-source multi-modal continual learning datasets, namely UESTC-MMEA-CL [33] and ARIC [34], consistently manifesting enhanced accuracy across diverse continual learning tasks.
- Moreover, we employ the decision attribution stability metric [5] to rigorously assess the interpretability of con-

*Authors contributed equally.
†Corresponding author.

tinual learner training implemented with frozen feature augmentation, empirically substantiating that our proposed feature augmentation strategy preserves the inherent interpretability of the learning framework.

2. Related Work

In this session, we review some related works about few-shot class-incremental learning and data augmentation.

2.1. Few-shot Class-Incremental Learning

In FSCIL, the training data in each incremental session is limited, worsening catastrophic forgetting and overfitting [29]. To mitigate these issues, recent works [36, 39] have emphasized representation learning, aiming to enhance the separation between base classes and reserve sufficient representation space for future novel classes. For instance, to ensure that the knowledge of future classes can be easily incorporated into the current model, FACT [39] reserves embedding space for future classes by assigning virtual prototypes. To further enhance inter-class separability and minimize disruption to previously learned knowledge, C-FSCIL [12] employs a quasi-orthogonal alignment of class prototypes. Building on these works, OrCo [1] introduces a novel framework that maximizes the margins between different classes by aligning class features with orthogonal pseudo-targets. In the field of multi-modal FSCIL, DIY-FSCIL [3] leverages cross-modal pretraining, weight generation, and knowledge distillation to efficiently learn new classes from minimal sketch examples while preventing catastrophic forgetting, whereas CPE-CLIP [8] introduces a parameter-efficient prompt-based framework that adapts CLIP [23] through learnable prompts to mitigate forgetting and enhance cross-task knowledge transfer. Our work targets to propose a feature augmentation strategy for FSCIL and validate it on both single- and multi-modal datasets.

2.2. Data augmentation

A common approach to enhance the performance in the few-shot environment is data augmentation [27]. MixUp [37] generates synthetic training samples by linearly interpolating between pairs of input data and their corresponding labels. Meanwhile, TrivialAugment [21] applies randomly selected fixed-intensity operations to each image to expand the dataset, whereas RandAugment [7] randomly selects augmentation operations and parameters. In addition, some recent studies have been proposed for feature augmentation [2, 9, 16]. For instance, FroFA [2] applies frozen feature augmentation to promote the performance of few-shot image classification. Inspired by this, our work incorporates feature augmentation into the FSCIL framework.

3. Feature Augmented FSCIL

3.1. Preliminaries

FSCIL Setting. In FSCIL, the continual learner is trained on a series of incremental sessions $\{\mathcal{D}^0, \mathcal{D}^1, \cdots, \mathcal{D}^T\}$, where $\mathcal{D}^\tau = \{(\mathbf{X}_{\tau,i}, y_{\tau,i})\}_{i=1}^{|\mathcal{D}^\tau|}$ represents the training set from session τ. $\mathbf{X}_{\tau,i}$ denotes the sample corresponding to the label $y_{\tau,i}$, which is usually a color image. Here, the first session dataset $\mathcal{D}^0$, also referred to as base dataset, consists of $|\mathcal{C}^0|$ classes, with each class $c \in \mathcal{C}^0$ containing sufficient training data. In each following incremental session ($\tau > 0$), only a limited number of labeled samples are available, with $|\mathcal{C}^\tau|$ classes, each containing K samples. The label set $\mathcal{C}^\tau$ from session τ has no overlap with that from any other sessions, *i.e.*, $\mathcal{C}^\tau \cap \mathcal{C}^{\tau'} = \emptyset, \forall \tau' \neq \tau$, where $\mathcal{C}^{\tau'}$ represents the label set of session τ'. After completing the τ-th session, the model's preformance is evaluated on test set from the current (τ-th) seesion and all previously encountered sessions ($< \tau$).

3.2. Frozen Feature Augmentation

During the few-shot incremental learning sessions of FSCIL, the scarcity of training data increases the risk of overfitting, which in turn degrades overall performance. This work proposes a novel frozen feature augmentation method, which enhances the diversity of few-shot class data by augmenting frozen features extracted from the pre-trained model, thereby effectively mitigating overfitting.

Specifically, for a certain sample $\mathbf{X}_{\tau,i}$ in session τ, we firstly employ the frozen pre-trained Vision Transformer (ViT) $f(\cdot)$ to obtain its frozen feature $\mathbf{Z}_{\tau,i} = f(\mathbf{X}_{\tau,i})$. For convenience, the subscript of $\mathbf{Z}_{\tau,i}$ is removed in the latter part of this section. The frozen feature $\mathbf{Z} = [\mathbf{z}_{cls}; \mathbf{z}_1; \mathbf{z}_2; \cdots; \mathbf{z}_N] \in \mathbb{R}^{(N+1)\times D}$ and $\mathbf{z}_{cls} \in \mathbb{R}^{1\times D}$ represents the class token utilized for classification. The tokens are selected to construct $\mathbf{Z}_{patch} = [\mathbf{z}_1, \mathbf{z}_2, \cdots \mathbf{z}_N] \in \mathbb{R}^{N\times D}$ corresponding to the image patches for feature augmentation. Then, we reshape $\mathbf{Z}_{patch}$ into a three-dimensional representation $\mathbf{Z}^*_{patch} \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times D}$ and apply Min-Max Normalization to each of the D feature channels, mapping each element to the range $\mathbb{I} = [0, 1]$:

$$\tilde{\mathbf{Z}}^*_{patch} = \frac{\mathbf{Z}^*_{patch} - z_{\min}}{z_{\max} - z_{\min}} \tag{1}$$

where $z_{\max}$ and $z_{\min}$ are the per-channel maximum and minimum, respectively. Subsequently, a distinct random value $v \sim \mathcal{U}(-\delta, \delta)$ is added to each of the D feature channels, where $\mathcal{U}$ denotes a uniform distribution. This process resembles brightness augmentation at the image level. We then map the augmented result back to the original feature value range to get the final representation:

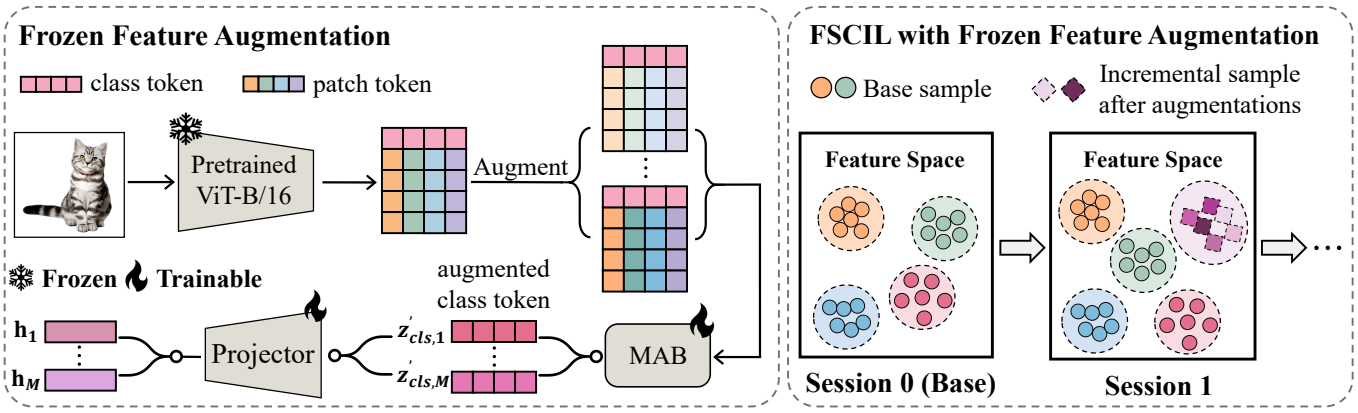


Figure 1. **FSCIL with Frozen Feature Augmentation. The left** illustrates the workflow of frozen feature augmentation. We extract the frozen feature using pre-trained ViT and augment image patch tokens. These augmented patch tokens are concatenated with the original class token and fed into the Multi-head Attention Block (MAB), producing the augmented class tokens $\{\mathbf{z}'_{cls,1}, \cdots, \mathbf{z}'_{cls,M}\}$. The augmented class tokens are projected to low-dimensional features $\{\mathbf{h}_1, \cdots, \mathbf{h}_M\}$ for classification. **The right** shows the process of using the frozen feature augmentation in the FSCIL task.

$$\hat{\mathbf{Z}}^*_{patch} = a\left(\tilde{\mathbf{Z}}^*_{patch}\right) \cdot (z_{\max} - z_{\min}) + z_{\min} \tag{2}$$

where $a(\cdot)$ represents the brightness operation. The result is then reshaped back into a two-dimensional representation, denoted as $\mathbf{Z}'_{patch} \in \mathbb{R}^{N \times D}$, and concatenated with the original class token $\mathbf{z}_{cls}$. This concatenated features are subsequently fed into the Multi-head Attention Block (MAB) $h(\cdot)$ to generate the final output:

$$\mathbf{z}'_{cls} = h\left([\mathbf{z}_{cls}; \mathbf{Z}'_{patch}]\right)[0, :] \in \mathbb{R}^D \tag{3}$$

where $[\cdot;\cdot]$ denotes the row-wise concatenation. MAB can fuse the augmented patch tokens $\mathbf{Z}'_{patch}$ into the original class token $\mathbf{z}_{cls}$ to obtain an augmented class token $\mathbf{z}'_{cls}$.

For the same few-shot class sample $\mathbf{X}$ from session τ ($\tau > 0$), we apply frozen feature augmentation to generate M augmented features $\{\mathbf{Z}'_1, \mathbf{Z}'_2, \cdots \mathbf{Z}'_M\}$, which are subsequently used to obtain M augmented class tokens $\left\{\mathbf{z}'_{cls,1}, \mathbf{z}'_{cls,2}, \cdots, \mathbf{z}'_{cls,M}\right\}$. Finally, an MLP projector $g(\cdot)$ will be trained to map these augmented class tokens from the original high-dimensional feature space to a lower-dimensional feature space. Frozen feature augmentation effectively enhances the data diversity of the few-shot class, thereby alleviating the overfitting in few-shot incremental learning sessions. This is one of the main innovations of this work.

3.3. FSCIL with Frozen Feature Augmentation

FSCIL aims to continually learn novel classes based on only few training samples. OrCo [1], as a typical FSCIL approach, introduces the OrCo framework to constrain the alignment of class features with pseudo-target vector, achieving SOTA performance. In this work, we will integrate our frozen feature augmentation method with the OrCo framework. The OrCo loss comprises three components: perturbed supervised contrastive loss, cross-entropy loss, and orthogonality loss.

Perturbed Supervised Contrastive Loss. To maximize the margins between distinct classes in the representation space, a set of mutually orthogonal vectors, referred to as pseudo-target vectors $\mathcal{T} = \{\mathbf{t}_c\}_{c=1}^{|\mathcal{T}|}$, $\mathbf{t}_c \in \mathbb{R}^d$, is first generated and then fixed, where $|\mathcal{T}|$ equals the total number of classes. Each class will be assigned a pseudo-target vector, which serves as class prototype for classification. Subsequently, a uniform perturbation is applied to these pseudo-target vectors, generating a series of perturbed pseudo-target vectors:

$$\tilde{\mathcal{T}} = \{\tilde{\mathbf{t}}_c\}_{c=1}^{|\tilde{\mathcal{T}}|}, \ \tilde{\mathbf{t}}_c = \mathbf{t}_c + \mathcal{U}(-\lambda, \lambda) \tag{4}$$

where $\mathcal{U}$ represents the uniform distribution and λ denotes the sampling boundary. In the perturbed supervised contrastive loss, the positive set includes not only other sample's features from the same class, but also the pseudo-target vector assigned to that class, as well as its perturbation. The negative set comprises sample features from other classes and other pseudo-target vectors. When incorporating frozen feature augmentation, different augmented features of the same sample will be included in the positive set, while the negative set will contain augmented features of samples from other classes.

Cross-entropy Loss. The cross-entropy loss is used to align all augmented features $\mathbf{h}_i = g\left(\mathbf{z}'_{cls,i}\right)$, $i \in [1, M]$ of a sample belonging to class $c \in \mathcal{C}^\tau$ with its assigned pseudo-target vector $\mathbf{t}_c$ in session τ.

Orthogonality Loss. During session τ, the purpose of the

orthogonality loss is to introduce implicit geometric constraints on class mean $\boldsymbol{\mu}_k = \frac{1}{N_k \times M} \sum_{j=1}^{N_k} \sum_{i=1}^{M} g\left(\mathbf{z}'_{cls,ji}\right)$ (N_k represents the number of samples of class k) and pseudo-target vectors $\hat{\mathcal{T}} = \{\mathbf{t}_c | c = 1, 2, \cdots |\mathcal{T}|, c \neq k\}$, *i.e.*, to ensure that each class mean and pseudo-target vectors remain orthogonal in the feature space, thereby reducing interference between different classes. The orthogonality loss takes the following form:

$$\mathcal{L}_{ORTH}\left(\boldsymbol{\mu}_k\right) = \frac{1}{\left|\hat{\mathcal{T}}\right|} \log \sum_{c=1}^{|\hat{\mathcal{T}}|} e^{\boldsymbol{\mu}_k^T \mathbf{t}_c / \tau_o} \tag{5}$$

where τ_o is the temperature parameter. In practice, we impose this constraint on all the augmented features within the current training batch.

To the end, the OrCo loss is a combination of the above three losses:

$$\mathcal{L}_{OrCo} = \mathcal{L}_{PSCL} + \mathcal{L}_{CE} + \mathcal{L}_{ORTH} \tag{6}$$

Here, $\mathcal{L}_{PSCL}$ represents the perturbed supervised contrastive loss and $\mathcal{L}_{CE}$ denotes the cross-entropy loss. Optimizing the OrCo loss facilitates the expansion of inter-class margins in the feature space, thereby reducing class confusion.

3.4. Inference

During inference, we first utilize the backbone (*i.e.*, the frozen pre-trained ViT) and the Multi-head Attention Block (MAB) to extract the class token $\mathbf{z}_{cls}$ corresponding to the test example $\mathbf{X}$. Next, the class token $\mathbf{z}_{cls}$ is fed into the MLP projector $g(\cdot)$ to obtain a d-dimensional feature $\mathbf{h} = g(\mathbf{z}_{cls})$. We compute the cosine similarity between $\mathbf{h}$ and the pseudo-target vectors $\mathcal{T} = \{\mathbf{t}_c\}_{c=1}^{|\mathcal{T}|}$ assigned to each class. The output class is corresponding to the largest cosine similarity:

$$\hat{c} = \underset{c \in \{\mathcal{C}^\tau\}_{\tau=0}^{T}}{\arg\max} \cos\left(\mathbf{h}, \mathbf{t}_c\right) \tag{7}$$

Here, $\{\mathcal{C}^\tau\}_{\tau=0}^{T}$ denotes the set of all classes, and $\cos(\cdot, \cdot)$ refers to cosine similarity.

4. Decision Attribution Stability Analysis

Frozen feature augmentation perturbs the original image features; therefore, we further investigate whether this perturbation affects the interpretability of the continual learners. Previous work [5] proposed two metrics based on the Shapley value, SHAPC-Mean and SHAPC-Var, to measure the continual learner's decision attribution stability. The two metrics first calculate the SHAP value for each feature component of a sample to quantify the importance of that component. Then, they assess the continual learner's decision attribution stability by evaluating the consistency of decision attribution from the same sample across different tasks. In this study, we leverage the two metrics to evaluate the impact of frozen feature augmentation on continual learner's interpretability.

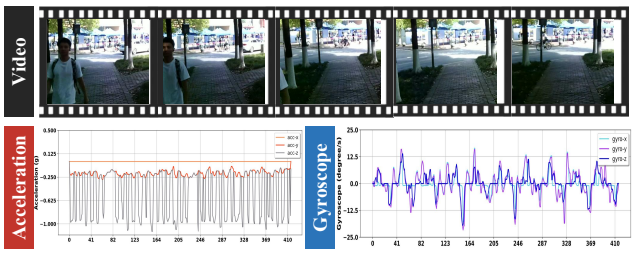


Figure 2. An example from UESTC-MMEA-CL.

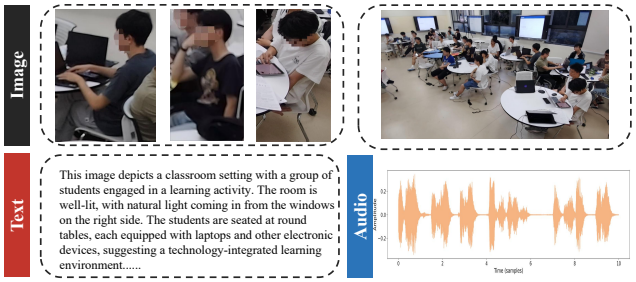


Figure 3. An example from ARIC. The image modality includes both the behavioral instance image and the corresponding classroom scene image. A description of the classroom, along with the audio comprising a 5-second clip, is used for the other two modalities.

5. Experiment

5.1. Datasets and Evaluation Metric

Datasets. The evaluation is conducted on two FSCIL benchmark dataset and two multi-modal datasets, *i.e.*, CIFAR-100 [14], mini-ImageNet [25], UESTC-MMEA-CL [33] and ARIC [34]. For CIFAR-100 and mini-ImageNet, following previous work [1], 60 object classes are set for base classes and 40 object classes are set for incremental classes with 8 sessions (5-way, 5-shot). UESTC-MMEA-CL is a multi-modal continual learning dataset, which contains three data modalities: video, accelerometer data, and gyroscope data, consisting of 32 classes of daily activities. An example is shown in Fig. 2. This dataset is divided into 16 base classes and 16 incremental classes structured in a 4-way, 5-shot FSCIL task. The ARIC dataset, which is derived from real classroom surveillance, encompasses 32 classroom activities across three modalities: image, text, and audio. Fig. 3 shows an example of ARIC. For this datasest, 20 classes are randomly selected as the base classes, with the remaining 12 classes being allocated into 4 sessions, each session consisting of a 3-way, 5-shot

Table 1. Comparison to prior works across the base and incremental sessions on **CIFAR-100** for 5-way, 5-shot setting. All the results are presented in the form of percentages (%)

Method	Base Acc	Session-wise Harmonic Mean (a_{hm}^{τ}) ↑								$\tilde{a}_{hm}$
		1	2	3	4	5	6	7	8	
IW [22]	78.58	42.31	44.79	41.28	40.26	38.88	39.75	39.40	38.22	40.61
CEC [36]	73.07	41.97	39.19	34.25	33.26	33.84	33.47	32.90	32.16	35.13
C-FSCIL [12]	77.47	31.47	28.33	26.68	23.11	24.85	24.29	23.04	23.49	25.66
LIMIT [40]	73.81	39.98	36.98	33.25	32.64	32.28	32.66	31.96	31.40	33.89
FACT [39]	74.60	42.85	38.21	33.02	32.34	31.70	34.15	33.37	33.61	34.91
BiDist [38]	69.68	47.87	42.68	40.19	37.01	35.42	33.80	33.25	30.94	37.65
NC-FSCIL [35]	82.52	56.66	54.41	49.70	45.26	44.36	46.65	44.21	41.98	47.90
OrCo [1]	80.08	72.12	63.92	56.90	55.21	53.37	54.04	51.78	49.56	57.11
OrCo-ViT	87.00	79.28	73.76	65.66	64.11	61.64	61.49	58.71	56.65	65.16
FAOrCo-ViT	**87.00**	**79.81**	**74.71**	**65.97**	**65.52**	**62.28**	**62.35**	**60.20**	**58.40**	**66.15**

Table 2. Comparison of different methods across base and incremental sessions on **mini-ImageNet** for 5-way, 5-shot setting.

Method	Base Acc	Session-wise Harmonic Mean (a_{hm}^{τ}) ↑								$\tilde{a}_{hm}$
		1	2	3	4	5	6	7	8	
IW [22]	83.10	49.49	45.09	45.98	46.30	44.67	42.48	43.26	45.65	45.36
FACT [39]	75.78	27.20	27.84	27.94	25.17	22.46	20.54	20.88	21.25	24.16
CEC [36]	72.17	31.91	31.84	30.98	30.74	28.14	26.78	26.96	27.42	29.35
C-FSCIL [12]	76.60	9.74	20.53	28.68	31.91	34.85	35.05	37.72	37.92	29.55
LIMIT [40]	73.27	40.34	33.58	31.81	31.74	29.32	29.11	29.57	30.28	31.97
LCwoF [15]	64.45	41.24	38.96	39.08	38.67	36.75	35.47	34.71	35.02	37.49
BiDist [38]	74.67	42.42	43.86	43.87	40.34	38.97	38.01	36.85	38.47	40.35
NC-FSCIL [35]	84.37	62.34	61.04	55.93	53.13	49.68	47.08	46.22	45.57	52.62
OrCo [1]	83.30	68.71	63.87	60.94	57.98	55.27	52.41	52.68	53.12	58.12
OrCo-ViT	96.05	94.42	90.96	89.93	89.76	88.49	87.14	86.68	86.81	89.27
FAOrCo-ViT	**96.05**	**94.53**	**92.03**	**90.80**	**90.74**	**89.29**	**87.79**	**87.35**	**87.61**	**90.02**

FSCIL task.

Evaluation Protocol. In this work, we use the harmonic mean to evaluate the model's classification accuracy during continual learning, especially in the few-shot incremental learning sessions. The harmonic mean is computed by combining the base class accuracy and incremental session accuracy: $a_{hm}^{\tau} = (2 \times a_0 \times a_{1:\tau}) / (a_0 + a_{1:\tau})$; where a_0 denotes the accuracy of the model on base classes after learning the τ-th session, and $a_{1:\tau}$ denotes the average accuracy across the first to τ-th few-shot incremental sessions after learning the τ-th session. Furthermore, we define a_{hm}^{T} as the harmonic mean calculated after the model completes learning the final session, and $\tilde{a}_{hm} = \frac{1}{T} \sum_{\tau=1}^{T} a_{hm}^{\tau}$ as the average harmonic mean across all incremental sessions.

5.2. Implementation Details

For CIFAR-100 and mini-ImageNet, we employ the self-supervised pre-trained Vision Transformer (DINO-1K [6]) as the backbone. For UESTC-MMEA-CL, we uniformly sample 8 frames from each video and use DINO-1K to extract features from each frame and use the average feature of the 8 frames. STFT is applied to convert accelerometer and gyroscope data into spectrograms, from which the features are extracted using Tiny-SSAST [10]. For ARIC, DINO-1K is used for feature extraction from image, RoBERTa [18] is used for feature extraction from text description, and ESResNeXt [11] is used for audio feature extraction. The final feature is the concatenation of the features from three modalities. Stochastic Gradient Descent (SGD) with momentum is employed to train the Multi-head Attention Block (MAB) and the MLP projector. The training epochs is 20 in the base session, and 100 epochs in each few-shot incremental session. A cosine learning rate schedule with a maximum learning rate of 0.1 is used. For CIFAR-100, the number of augmentations $M = 4$ with the augmentation intensity $\delta = 0.1$, while for mini-ImageNet, $M = 5$ with the intensity of $\delta = 0.05$. We apply augmentation solely on the visual features for multi-modal data. For UESTC-MMEA-CL, M is set to 2, with an intensity δ of 0.1. For ARIC, M is set to 4, with an intensity δ of 0.1. During learning new session, the 5 saved exemplars per learned class are replayed.

Table 3. Results on Multi-modal datasets. All the results are presented in the form of percentages (%)

Method	UESTC-MMEA-CL		ARIC	
	$a_{hm}^{T}(\uparrow)$	$\tilde{a}_{hm}(\uparrow)$	$a_{hm}^{T}(\uparrow)$	$\tilde{a}_{hm}(\uparrow)$
OrCo-ViT	77.97	77.06	40.31	41.88
FAOrCo-ViT	**79.11**	**78.26**	**41.55**	**42.98**

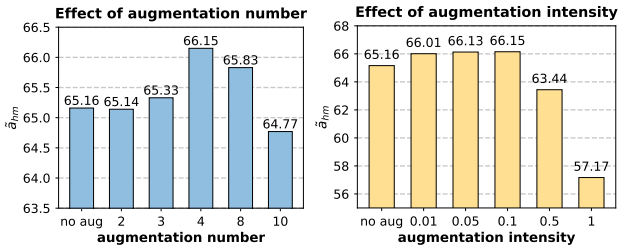
5.3. Comparison Results

Results on Regular Datasets. This section presents a comparison of the proposed method with recent state-of-the-art FSCIL methods. The experimental results on CIFAR-100 are reported in Table 1. OrCo-ViT refers to replacing the backbone in the original OrCo [1] framework with DINO-1K, incorporating the MAB introduced in this work. It serves as a baseline FSCIL method without frozen feature augmentation. On CIFAR100, our method achieves the highest harmonic mean in the final session a_{hm}^{T}, representing a 1.6% improvement over OrCo-ViT. Additionally, our method maintains the highest performance in terms of the average harmonic mean $\tilde{a}_{hm}$. Table 2 presents the experimental results on mini-ImageNet, where our method demonstrates consistent effectiveness, achieving the highest a_{hm}^{T} (87.61%) and $\tilde{a}_{hm}$ (90.02%) among all compared methods. Notably, the proposed method consistently exhibits substantial improvement across all incremental sessions on both CIFAR-100 and mini-ImageNet.

Results on Multi-modal Datasets. Furthermore, we validate our method on two multi-modal datasets: UESTC-MMEA-CL and ARIC. The results are presented in Tabel 3. For UESTC-MMEA-CL, our method achieves a 1.1% improvement in a_{hm}^{T} and a 1.2% improvement in the average harmonic mean $\tilde{a}_{hm}$ compared with OrCo-ViT. For ARIC, our method still has a 1.24% improvement in a_{hm}^{T} and a 1.1% improvement in $\tilde{a}_{hm}$. The results further demonstrate that even when frozen feature augmentation is applied solely to the features of the image modality, the diversity of the training samples in the few-shot class can still be effectively enhanced, thereby improving the performance of continual learning.

5.4. Parameter Analysis and Model Validation

Comparison of different augmaneted number. We first investigate how the augmented number of features affects FSCIL results. Fig. 4a shows the histograms of the average harmonic mean $\tilde{a}_{hm}$ against M, where M is set to 2, 3, 4, 8, and 10. The intensity δ is consistently fixed at 0.1. This result reveals that when the number of augmentations is low, the diversity of sample features is insufficient to effectively mitigate overfitting. On the other hand, excessively increasing the number of augmentations may cause the training data distribution to deviate from the original data distribution, leading to a decrease in model performance.



(a) (b)

Figure 4. Ablation results on CIFAR-100 showing the effect of augmentation number (left) and augmentation intensity (right) on the average harmonic mean $\tilde{a}_{hm}$

Table 4. Comparison of experimental results for different augmentations on CIFAR-100.

Augmentation	CIFAR-100	
	$a_{hm}^{T}(\uparrow)$	$\tilde{a}_{hm}(\uparrow)$
None	56.65	65.16
Contrast	53.92	65.15
Sharpness	57.77	65.66
Brightness (Ours)	**58.40**	**66.15**

Comparison of different augmentation intensity. We believe that the augmentation intensity δ is a critical parameter influencing augmentation performance; hence, we conduct an in-depth investigation. The histograms of the average harmonic mean $\tilde{a}_{hm}$, for δ values of 0.01, 0.05, 0.1, 0.5, and 1.0 are presented in Fig. 4b. The results indicate that augmentation performance declines significantly when δ is excessively high. This suggests that an excessively high degree of augmentation can cause the augmented sample features to deviate significantly from the original feature distribution, resulting in the loss of critical information necessary for classification.

Table 5. Comparison of experimental results for different framework on CIFAR-100.

Method	CIFAR-100	
	$a_{hm}^{T}(\uparrow)$	$\tilde{a}_{hm}(\uparrow)$
OrCo	49.56	57.11
OrCo-SwinT	64.59	72.81
FAOrCo-SwinT	**65.55**	**73.51**

Comparison of different augmentations. In addition to the augmentation scheme utilized in this study, we further compare our scheme with other augmentation schemes. The results are presented in Table 4. The item "None" indicates the absence of feature augmentation but already en-

Table 6. Results of decision stability evaluation on four datasets. All the results are presented in the form of percentages (%)

Method	CIFAR-100		mini-ImageNet		UESTC-MMEA-CL		ARIC	
	SHAPC		SHAPC		SHAPC		SHAPC	
	Mean(↑)	Var(↓)	Mean(↑)	Var(↓)	Mean(↑)	Var(↓)	Mean(↑)	Var(↓)
OrCo-ViT	45.21	13.25	60.66	**10.61**	**44.19**	17.08	**38.83**	**8.30**
FAOrCo-ViT	**45.90**	**13.19**	**60.74**	10.90	44.12	**17.06**	38.12	8.39

compasses standard image augmentations (such as color perturbation and random cropping). Other items signify further frozen feature augmentations. "Contrast" refers to randomly scaling each of the D feature channels by a factor $v \sim \mathcal{U}(1/\delta, \delta)$. In this context, δ is set to 2.0. "Sharpness" involves applying a two-dimensional convolution with a 3×3 smoothing filter to each of the D feature channels after transforming each into a $\sqrt{N} \times \sqrt{N}$ tensor. The original features are then combined with the resulting smoothed features using a randomly sampled blending factor $v \sim \mathcal{U}(0, \delta)$, where δ is set to 1.5. "Brightness" is our augmentation method, where a random value $v \sim \mathcal{U}(-\delta, \delta)$ is added to each of the D feature channels. The augmentation number M is consistently set to 4. The experimental results indicate that our brightness augmentation surpasses tother two methods, *i.e.*, "Contrast" and "Sharpness".

Comparison across different framework. Our method is further validated using the Swin Transformer [19], with the experimental results on CIFAR-100 presented in Table 5. The results show that employing the frozen feature augmentation (FAOrCo-SwinT) improves a_{hm}^T by 1% and $\tilde{a}_{hm}^T$ by 0.7% compared to the baseline method (OrCo-SwinT). These findings demonstrate effectiveness of our method across different network architectures.

5.5. Decision Stability Analysis

We evaluate the interpretability of the learned continual learner with feature augmentation (*i.e.*, FAOrCo-ViT) and without feature augmentation (*i.e.*, OrCo-ViT) through decision attribution analysis outlined in Section 4. The experimental results are presented in Table 6. The results demonstrate comparable performance between FAOrCo-ViTand OrCo-ViT across all datasets, with minimal differences in both metrics (typically $< 0.8\%$ for SHAPC-Mean and $< 0.3\%$ for SHAPC-Var). These marginal differences indicate that our feature augmentation scheme does not compromise the interpretability of the continual learner.

6. Conclusion

This work proposes a frozen feature augmentation method for FSCIL, which enhances FSCIL's performance on both regular and multi-modal continual learning datasets. Through decision attribution analysis, we demonstrate that this feature augmentation scheme does not compromise the interpretability of the continual learner. Our future work will focus on exploring more efficient frozen feature augmentation schemes in the context of FSCIL.

7. Acknowledgement

This work is supported by the National Science and Technology Major Project (2021ZD0112001), and National Natural Science Foundation of China (No. 62171111).

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. 1, 2, 3, 4, 5, 6

[2] Andreas Bär, Neil Houlsby, Mostafa Dehghani, and Manoj Kumar. Frozen feature augmentation for few-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16046–16057, 2024. 2

[3] Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Doodle it yourself: Class incremental learning by drawing a few sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2293–2302, 2022. 2

[4] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020. 1

[5] Yusong Cai, Shimou Ling, Liang Zhang, Lili Pan, and Hongliang Li. Is our continual learner reliable? investigating its decision attribution stability through shap value consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5568–5575, 2024. 1, 4

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation

strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 2

[8] Marco D'Alessandro, Alberto Alonso, Enrique Calabrés, and Mikel Galar. Multimodal parameter-efficient few-shot class incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3393–3403, 2023. 2

[9] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017. 2

[10] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10699–10709, 2022. 5

[11] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 5

[12] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9057–9067, 2022. 1, 2, 5

[13] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 1

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[15] Anna Kukleva, Hilde Kuehne, and Bernt Schiele. Generalized and incremental few-shot learning by explicit learning and calibration without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9020–9029, 2021. 5

[16] Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and William Campbell. A closer look at feature space data augmentation for few-shot intent classification. *arXiv preprint arXiv:1910.04176*, 2019. 2

[17] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 1

[18] Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. Roberta: A robustly optimized bert pretraining approach. arxiv [preprint](2019). *arXiv preprint arXiv:1907.11692*, 1907. 5

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[20] Pratik Mazumder and Pravendra Singh. Mitigate forgetting in few-shot class-incremental learning using different image views. *Neural Networks*, 165:999–1009, 2023. 1

[21] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 774–782, 2021. 2

[22] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. 5

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[24] 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. 1

[25] 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. 4

[26] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 1

[27] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 2

[28] 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. 1

[29] 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. 2

[30] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324, 2024. 1

[31] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[32] 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. 1

[33] Linfeng Xu, Qingbo Wu, Lili Pan, Fanman Meng, Hongliang Li, Chiyuan He, Hanxin Wang, Shaoxu Cheng, and Yu

Dai. Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning. *IEEE Transactions on Multimedia*, 2023. 1, 4

[34] Linfeng Xu, Fanman Meng, Qingbo Wu, Lili Pan, Heqian Qiu, Lanxiao Wang, Kailong Chen, Kanglei Geng, Yilei Qian, Haojie Wang, et al. Aric: An activity recognition dataset in classroom surveillance images. *arXiv preprint arXiv:2410.12337*, 2024. 1, 4

[35] 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. 1, 5

[36] 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. 1, 2, 5

[37] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[38] 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. 5

[39] 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. 1, 2, 5

[40] Da-Wei Zhou, Han-Jia Ye, Liang Ma, Di Xie, Shiliang Pu, and De-Chuan Zhan. Few-shot class-incremental learning by sampling multi-phase tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12816–12831, 2022. 1, 5

[41] Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: A survey. *arXiv preprint arXiv:2401.16386*, 2024. 1

[42] Mingli Zhu, Zihao Zhu, Sihong Chen, Chen Chen, and Baoyuan Wu. Enhanced few-shot class-incremental learning via ensemble models. *arXiv preprint arXiv:2401.07208*, 2024. 1