Rethinking Few-shot Class-incremental Learning: Learning from Yourself

Yu-Ming Tang[1,3], Yi-Xing Peng[1,3,5], Jingke Meng⋆[1,3], and Wei-Shi Zheng[1,2,3,4,5]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4] Guangdong Province Key Laboratory of Information Security Technology, Sun Yat-sen University, China
[5] Pazhou Laboratory (Huangpu), Guangzhou, Guangdong 510555, China
`{tangym9, pengyx23}@mail2.sysu.edu.cn`, `mengjke@gmail.com`, `wszheng@ieee.org`

Abstract. Few-shot class-incremental learning (FSCIL) aims to learn sequential classes with limited samples in a few-shot fashion. Inherited from the classical class-incremental learning setting, the popular benchmark of FSCIL uses averaged accuracy ($aAcc$) and last-task averaged accuracy ($lAcc$) as the evaluation metrics. However, we reveal that such evaluation metrics may not provide adequate emphasis on the novel class performance, and the continual learning ability of FSCIL methods could be ignored under this benchmark. In this work, as a complement to existing metrics, we offer a new metric called generalized average accuracy ($gAcc$) which is designed to provide an extra equitable evaluation by incorporating different perspectives of the performance under the guidance of a parameter α. We also present an overall metric in the form of the area under the curve (AUC) along the α. Under the guidance of $gAcc$, we release the potential of intermediate features of the vision transformers to boost the novel-class performance. Taking information from intermediate layers which are less class-specific and more generalizable, we manage to rectify the final features, leading to a more generalizable transformer-based FSCIL framework. Without complex network designs or cumbersome training procedures, our method outperforms existing FSCIL methods at $aAcc$ and $gAcc$ on three datasets. See codes at [1]

Keywords: Continual Learning · Few-shot Learning · Catastrophic Forgetting · Vision Transformer

1 Introduction

Although existing deep neural networks (DNNs) achieve great success in various real-world applications [17, 31, 32, 34, 43, 44, 54, 68], they suffer from *catastrophic*

⋆ Corresponding Author

[1] `https://github.com/iSEE-Laboratory/Revisting_FSCIL`

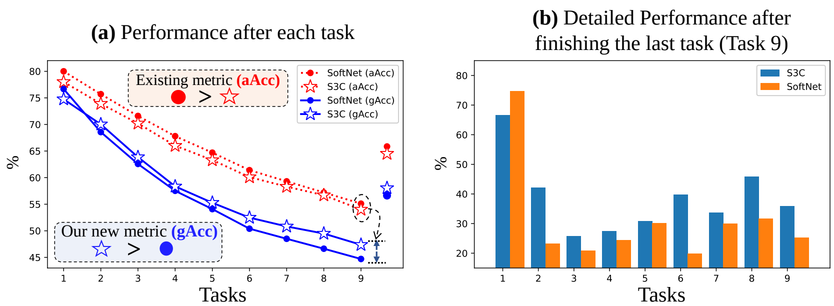

Fig. 1: Performance under different metrics of two recent methods [24, 25]. **(a)**: The dotted line denotes the *average accuracy (aAcc)* widely used in classical class-incremental learning. The solid line presents our proposed *generalized accuracy (gAcc)*. We also show the average of each point on the right side. (Blue: *gAcc*, Red: *aAcc*). **(b)**: The detailed accuracies of each task after training on the last task (task 8). It is evident that S3C exhibits superior performance when confronted with novel classes. The conventional metric *aAcc* fails to reflect this due to the domination by the base-class performance. Models are trained on the CIFAR-100 dataset. Best viewed in color.

forgetting when dealing with sequential data streams. The forgetting problem has attracted heavy attention in recent years [4,16,21,29,33,45,48,50,51,57,61,64,65, 71, 74]. In this work, we focus on a more realistic and challenging problem: few-shot class-incremental learning (FSCIL), in which after training on a task with sufficient data (**base** task), only a few samples are provided during incremental tasks (**novel** tasks). Since Tao *et al.* [52] proposed the FSCIL benchmark in 2020, a large number of algorithms have emerged to tackle this problem, including meta-learning based [8, 20, 70], structure based [2, 25, 52, 58, 63], feature space based [3, 6, 20, 24, 26, 59, 69] methods, *etc.*

However, almost all these methods are evaluated following the traditional class-incremental learning setting, *i.e.* using the average accuracy ($aAcc$) or the last-task averaged accuracy ($lAcc$) as the main metrics to critique the proposed method. Such accuracies are calculated by weighted averaging the base-task and novel-tasks accuracies with the weight proportional to the number of classes of each task. Considering the reality that the FSCIL benchmarks make the majority of the classes (about 50% to 60%) as the base task to establish a well-trained backbone for few-shot continual learning, the existing $aAcc$ or $lAcc$ metric is **dominated** by the base-class performance. In other words, current evaluation metrics can be largely boosted by only improving the performance of base classes. As shown in Fig. 1, a method like softNet [25] with higher base-class performance (see Fig. 1 (b)) usually gets higher $aAcc$ (see red lines in Fig. 1 (a)) regardless of the inferior performance on the novel classes. Due to the inadequacy to capture performance changes in novel classes, current metrics are biased toward the performance of base classes and have limitations in guiding the FSCIL community. To tackle this, we propose a more balanced metric called *generalized accuracy ($gAcc$)*. Parameterized by α ranging from 0 to 1, the $gAcc$ offers a way to evaluate a model with explicit emphasis on performance on novel classes. Furthermore, we can comprehensively evaluate the model by calculating the area under the curve (AUC) along the parameter α. It is worth mentioning that our intention is NOT to replace all previous metrics, but to provide a more fine-grained and balanced evaluation of FSCIL methods complementary to them.

We empirically find that FSCIL can benefit from the powerful vision transformer (ViT) [15] by analyzing the performance of novel classes. Specifically, although vanilla ViT performs poorly in terms of unseen novel classes, we reveal that the intermediate features of ViT are **less** specified on base classes. It means the ViT can learn abundant semantic information but the information is mistakenly dropped in deep layers when extracting features for novel classes.

Based on the observation, we propose a lightweight Feature Rectification (FR) module to rectify the output features of the ViT conditioned on the intermediate features. To achieve this, the FR module is supervised with two losses that work on two different levels, namely instance-level relation transfer and class-center relation transfer. The former transfers the relations between instances while the latter focuses on relations between instances and the class prototypes. We further enhance the FR module by leveraging hierarchical knowledge from multiple intermediate layers. Finally, our method boosts the novel-class accuracy and maintains a comparable performance of base classes.

Our contributions are summarized as follows. (1) We revisit the existing FSCIL benchmark and propose a novel metric called *generalized accuracy (gAcc)* for a more fine-grained and balanced evaluation. We have reimplemented 14 FSCIL methods with codes released and show that the newly introduced SOTAs do not necessarily perform better on *gAcc*. (2) Based on *gAcc*, we reveal the potential of vision transformers in FSCIL. We empirically find that although features from the intermediate layer exhibit sub-optimal performance for the base classes, their transferrable ability helps them to have a distinct advantage in handling novel classes. (3) A lightweight FR module is proposed to rectify the final features to be more transferrable. The FR module manages to make the output feature more generalizable with assistance from intermediate features. (4) We conduct extensive experiments on 3 datasets and evaluate our method under both *aAcc* and *gAcc*, showing the effectiveness and superiority of our method.

2 Related Work

Meta-learning based methods. Due to the limited samples of novel classes, some works [8, 20, 67, 70] propose to use meta-learning to improve the transferability of the model. Concretely, MetaFscil [8] proposed to simulate novel tasks by sampling from the base task. C-FSCIL [20] uses meta-learning to train the backbone, which maps images from different classes to quasi-orthogonal vectors. ALFSCIL [30] formulate the process into a bi-level optimization problem.

Optimization-based methods. There is a large group of works focusing on the optimization of the DNNs in FSCIL. CLOM [76] considers the behavior of the backbone under different margins of the cosine loss. NC-FSCIL [59] designs a pre-defined classifier for efficient FSCIL without any explicit regularizer. SPPR [73] proposes a prototype refinement strategy for stronger expression ability. FACT [69] assigns virtual prototypes in the current stage for future updates. Alice [42] instead focuses on the compactness and the variety of the feature space. Based on the classifier learning, S3C [24] proposes a self-supervised stochastic classifier to tackle the overfitting and forgetting problem. There are more works working on the feature space constraint [3, 6, 13, 26, 27, 56, 60], or knowledge distillation [7, 9–11, 14, 39, 66, 72].

Data-driven methods. There is also a branch of work focusing on the data side of this problem, which focuses on the data replay [1,35]. FSIL-GAN [1] employs a GAN to generate features to compensate for the scarcity of data. DF-replay [35] uses synthesized data generated by a generator to replay. PL-FSCIL [23] seeks prior knowledge from unlabeled data.

Structure-based methods. In addition to these works, a unique branch of works focuses on the network structure, like topology preserving [52], growing classifier [63], proposing a dynamic structure of the backbone [58], structure fusion [2, 22], sub-networks [25] and ensemble models [75].

Compared with existing works, our work rephrases the challenge of the FSCIL setting and first reveals the potential of the intermediate features of the ViT backbone. There is also an existing work [42] offering an alternative metric. They did not provide a fine-grained evaluation with various perspectives and have potential risks in stability and symmetry problems. Please find more detailed discussions about related works in the supplementary material.

3 Generalized Average Accuracy

3.1 Limitation of the Existing Metrics

Accuracy metric on incremental tasks. Consider a classification model $f(\cdot)$ that takes an image as input and outputs the class predictions. For an incremental problem, there are multiple tasks and multiple test sets. Let us define test set at task $\mathcal{T}_i$ as $\mathcal{D}^i_{test} = \{x^i_t, y^i_t\}^{n^i}_{t=1}$. Usually, the labeling space $\mathcal{Y}^i \triangleq \{y^i_t | (x^i_t, y^i_t) \in \mathcal{D}^i_{test}\}$ of every task does not have any intersection, *i.e.* $\forall i \neq j, \mathcal{Y}^i \bigcap \mathcal{Y}^j = \emptyset$. After trained on $\mathcal{T}_i$, the model can be denoted as $f_i(\cdot)$. The corresponding accuracy of model $f_i(\cdot)$ at task $\mathcal{T}_j$ $(j \leq i)$ can be denoted as $\mathcal{A}^j_i \triangleq \frac{1}{n^j} \sum_{t=1}^{n^j} \mathbb{1}(\zeta(f_i(x^j_t)) = y^j_t)$, where $\mathbb{1}(\cdot) = 1$ if the condition is true, otherwise 0. ζ is the argmax operation.

Average accuracy. It is worth mentioning that for class-incremental learning, evaluating the model on **all** the previous tasks is essential. To test the model on previous tasks, an overall test set that contains all previous test sets is constructed as $\mathcal{D}^{1\to i}_{test} = \bigcup_{m=1}^{i} \mathcal{D}^m_{test}$. After trained on task $\mathcal{T}_i$, the accuracy of model $f_i(\cdot)$ on the overall test set $\mathcal{D}^{1\to i}_{test}$ can be presented as $A^{1\to i}_i$. Once a n_t-task incremental training ends, the *average accuracy (aAcc)* can be calculated as:

$$aAcc = \frac{1}{n_t} \sum_{i=1}^{n_t} aAcc_i = \frac{1}{n_t} \sum_{i=1}^{n_t} A^{1\to i}_i, \tag{1}$$

where the $aAcc_i$ is the average accuracy at each task. We can also derive two additional metrics incidentally: last-task averaged accuracy $lAcc = aAcc_{n_t} = A^{1\to n_t}_{n_t}$, and task-wise average accuracy ($tAcc$):

$$tAcc = \frac{1}{n_t} \sum_{i=1}^{n_t} tAcc_i = \frac{1}{n_t} \sum_{i=1}^{n_t} \frac{1}{i} \sum_{j=1}^{i} A^j_i, \tag{2}$$

where the $tAcc_i$ is the task-wise accuracy at each task. Note that the $aAcc$ does not share the same value as $tAcc$ normally since the number of classes of tasks is usually not the same. This will be elaborated on later in detail.

Limitation. Since FSCIL usually sets the majority of the classes as the base task, existing metrics potentially neglect the performance changes of novel classes. To elaborate on this, we analyze the $aAcc$ as follows. Firstly, we denote the size of the labeling space of task $\mathcal{T}_i$ as $|\mathcal{Y}_i|$. In FSCIL, the number of classes of each novel task is usually the same, *i.e.* $|\mathcal{Y}_{novel}| = |\mathcal{Y}_2| = |\mathcal{Y}_3| = \cdots = |\mathcal{Y}_{n_t}| \ll |\mathcal{Y}_1|$. Assuming there are total n_t tasks, and task $\mathcal{T}_1$ is the *base task* with sufficient classes and samples (non-few-shot). The following tasks $\{\mathcal{T}_i\}_{i=2}^{n_t}$ are *novel tasks*, with $|\mathcal{Y}_{novel}|$ classes for each task and a few samples per class. Now, we can rewrite average accuracy on task $\mathcal{T}_i$ as:

$$aAcc_i = \frac{|\mathcal{Y}_1|A_i^1 + |\mathcal{Y}_{novel}|\sum_{j=2}^{i} A_i^j}{|\mathcal{Y}_1| + (i-1)|\mathcal{Y}_{novel}|} = \frac{\frac{|\mathcal{Y}_1|}{|\mathcal{Y}_{novel}|}A_i^1 + \sum_{j=2}^{i} A_i^j}{\frac{|\mathcal{Y}_1|}{|\mathcal{Y}_{novel}|} + (i-1)}. \tag{3}$$

From Eq. (3) we see that the value of $aAcc_i$ will be dominated by the base class accuracy A_i^1 if the ratio $\frac{|\mathcal{Y}_1|}{|\mathcal{Y}_{novel}|}$ is large. Unfortunately, the ratio is indeed large in the FSCIL settings. (*e.g.* $\frac{|\mathcal{Y}_1|}{|\mathcal{Y}_{novel}|} = 12$ for both CIFAR-100 and *mini*ImageNet dataset, 10 for CUB-200.) Under such a large ratio, the changes (both improvement and degrades) of novel classes cannot be effectively evaluated by $aAcc$.

3.2 Generalizing the Average Accuracy

To overcome the limitations of existing metrics, we now define the *generalized accuracy* ($gAcc$) at task $\mathcal{T}_i$ as:

$$gAcc_i(\alpha) = \frac{\alpha\frac{|\mathcal{Y}_1|}{|\mathcal{Y}_{novel}|}A_i^1 + \sum_{j=2}^{i} A_i^j}{\alpha\frac{|\mathcal{Y}_1|}{|\mathcal{Y}_{novel}|} + (i-1)}, \tag{4}$$

in which the $gAcc$ is parameterized by $\alpha \in [0, 1]$. At task $\mathcal{T}_i$, $gAcc_i(0)$ measures the accuracy of the novel tasks only. $gAcc_i(\frac{|\mathcal{Y}_{novel}|}{|\mathcal{Y}_1|})$ denotes the task-wise accuracy $tAcc_i$ at task $\mathcal{T}_i$ while $gAcc(1)$ represents the average accuracy $aAcc_i$ at task $\mathcal{T}_i$. In addition to these discrete values, we further generalize the α to any rational numbers in $[0, 1]$. Controlled by α in Eq. (4), when $\alpha \to 0$, $gAcc(\alpha)$ emphasizes to novel class while $\alpha \to 1$, $gAcc(\alpha)$ emphasizes the base class performance, *i.e.*, the larger the α, the greater the influence of the base class on the value of $gAcc(\alpha)$. To evaluate the performance of a model on a specific task $\mathcal{T}_i$, a standalone value is needed. Thus, we introduce the area under the curve (AUC) of $gAcc_i(\alpha)$ along the parameter α as a comprehensive metric:

$$gAcc_i = \int_0^1 gAcc_i(\alpha)\, d\alpha. \tag{5}$$

Similar to $aAcc$, for multiple tasks, we average the $gAcc_i$ for each task and get an overall metric $gAcc$ for the whole learning process:

$$gAcc = \frac{1}{n_t}\sum_{i}^{n_t} gAcc_i. \tag{6}$$

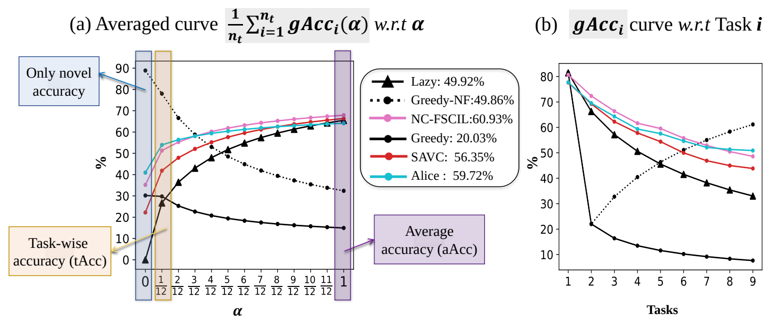

Fig. 2: FSCIL performances are shown in *generalized accuracy*. (a): The $gAcc$ curve (averaged across all n_t tasks) v.s. param α. (b): The $gAcc$ AUC of each task $\mathcal{T}_i$ (Eq. (5)). In the legend, we show the average AUC across tasks(Eq. (6)) of each method. We evaluate recent works including SAVC [49], NC [59], and Alice [42] on the *mini*ImageNet dataset. Some corner cases: **Lazy**: The model maintains base-task performance while refusing to learn anything from novel tasks. **Greedy**: Regardless of the previous knowledge, the model only greedily focuses on the current task. **Greedy-NF**: A **N**on-**F**orget version of 'Greedy'. See more about these cases in our supplementary material.

The proposed $gAcc$ provides different evaluation perspectives using different α so we believe such a metric can provide a more balanced evaluation of FSCIL.

3.3 Analysis

We further study the necessity of the proposed $gAcc$. We evaluate recent SOTAs in FSCIL including SAVC [49], Alice [42] and NC-FSCIL [59], and the results are in Fig. 2. We show that the $gAcc$ provides more fine-grained evaluation than the classical accuracy metric. For example, in the left part of Fig. 2, although the $gAcc(1)$, which is equal to the $aAcc$, of SAVC is close to that of NC-FSCIL, the area under the curve(AUC) of SAVC is much smaller than NC-FSCIL. This indicates that the novel class performance is much worse for SAVC, which cannot be discovered by conventional metrics.

Furthermore, to show the superiority of our metric, we consider three corner cases: Lazy, Greedy, and Greedy-NF. 'Lazy' stands for a type of algorithm that **only** focuses on maintaining **base-class** performance, *i.e.* $\forall i \geq j > 1, A_i^j = 0, A_i^1 = A_1^1$. By setting $A_1^1 = 85\%$, we can see that when $\alpha = 1$ (*i.e.* the traditional $aAcc$ metric), its performance is close to SOTA methods. Further, when considering popular forgetting metrics like performance dropping ($PD = A_i^j - A_j^j$) and Knowledge rate ($KR = A_i^j / A_j^j$), because the forgetting is zero, its PD is 0 and KR is 1. All these metrics indicate such 'Lazy' model is superior, which is contradictory to common sense. However, our proposed $gAcc$ reveals that its performance is far from satisfying by setting $\alpha < 1$ and its AUC is far smaller than other methods. In other words, using the traditional metrics will indicate algorithms like 'Lazy' achieving SOTA performance while such methods perform poorly evaluated by our proposed $gAcc$. The $gAcc$ can also discover other corner cases. 'Greedy' is to greedily learn the current knowledge (*i.e.* $A_1^1 = 85\%$, $A_i^j = 100\%$, when $i = j$ else 0). "Greedy" achieves low both $aAcc$ and $gAcc$. We further eliminate the forgetting problem from 'Greedy' and define 'Greedy-NF' on top of it: set accuracy A_i^j to 100% for $\forall j > 1, j < i$ (no forgetting at novel classes). We find that although 'Greedy-NF' performs

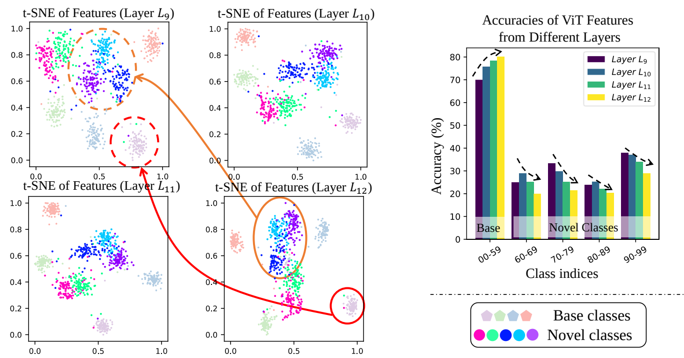

Fig. 3: t-SNE [37] visualization of *mini*ImageNet test set features from different layers of a ViT. **Left:** we choose layers L_9, L_{10}, L_{11}, L_{12}(final) and show both base classes and novel classes features. It is observed that shallow features are more dispersed than deep features. **Right:** Accuracies of different tasks on features from various layers. It is clear that intermediate features achieve better performance on novel classes and thus have better generalization ability. Best viewed in color.

well when majorly considering the novel-class performance ($\alpha \in \{0, \frac{1}{12} \frac{2}{12} \frac{3}{12}\}$), the overall performance (AUC) is not satisfying. Please refer to Sec. A1 in the supplementary material for more details about these corner cases.

By analyzing these cases, we show the superior properties of the proposed $gAcc$. Firstly, the $gAcc$ metric effectively distinguishes methods (*e.g.* 'Lazy') that only retain the base-task performance, thereby encouraging methods to learn from novel classes. Besides, $gAcc$ avoids overemphasis on the performance of base classes or novel classes (corner case 'Greedy-NF' is still far worse than normal methods in terms of AUC), providing a more balanced evaluation outcome. In other words, the proposed $gAcc$ metric provides objective and balanced results when evaluating FSCIL methods, eliminating 'shortcuts' in $aAcc$ that focus only on the performance of the base classes.

In conclusion, the $gAcc$ curve and the AUC are necessary and powerful tools to evaluate the FSCIL methods.

4 Learn from Yourself

4.1 Preliminary and The FSCIL Setting

For task $\mathcal{T}_i \in \{\mathcal{T}_i\}_{i=1}^T$, the training set can be denoted as $\mathcal{D}_{train}^i = \{x_t^i, y_t^i\}_{t=1}^{m^i}$. For FSCIL settings, the base task $\mathcal{T}_1$ contains $|\mathcal{Y}_1|$ classes and sufficient samples for training. For $\forall i > 1$, there are $|\mathcal{Y}_{novel}|$ classes for each task and k samples per class. Typically, $|\mathcal{Y}_{novel}|$ is set to 5, and $k = 5$, *i.e.* a 5-way 5-shot setting. We rephrase the classification model $f(\cdot)$ in Sec. 3 as $f(\cdot) = \Phi(f_B(\cdot))$ where $\Phi(\cdot)$ is the classifier and $f_B(\cdot)$ is the backbone. We follow previous works to use a cosine classifier: $\Phi(X) = P^\top \overline{X}$, where $P \in \mathbb{R}^{d \times |\mathcal{Y}|}$ and $\overline{X} \in \mathbb{R}^{d \times 1}$ are the L2-normalized inputs, and d is the feature dimension of $f_B(\cdot)$, $|\mathcal{Y}| = \sum_i |\mathcal{Y}_i|$ is the number of classes seen so far. In this work, we employ the Vision Transformer(ViT) [15] in the FSCIL setting. For a ViT with N_L layers $\{L_l\}_{l=1}^{N_L}$, given a input image x, the output can be denoted as $X_{N_L} = L_{N_L}(L_{N_L-1}(\cdots L_1(x))$.

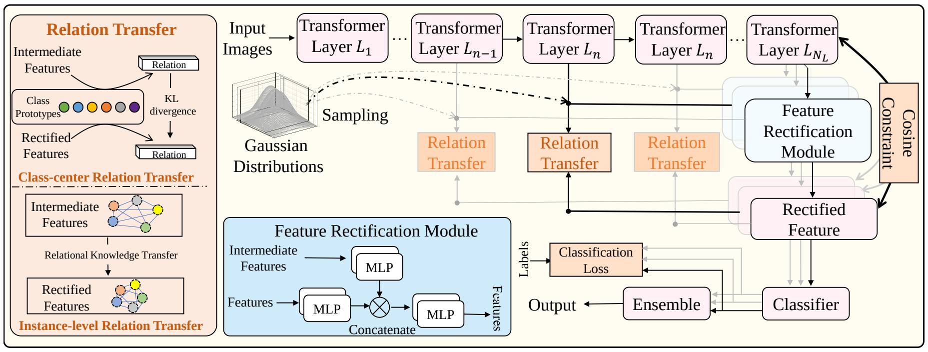

Fig. 4: The overall framework of our method. The Feature Rectification (FR) module takes the final and the intermediate features as input and outputs the rectified features for classification. The relation transfer losses are designed to convey valuable information from intermediate features to the rectified ones. The details of these two parts are shown in the left and lower parts of the figure. Further, a cosine constraint is proposed to maintain the base class performance and a classification loss is employed to adapt to the novel classes. Best viewed in color.

Baseline. Recent works [3, 20, 24, 38, 47, 49, 59, 63, 76] suggest a training-frozen scheme, which trains the backbone $f_B(\cdot)$ and the classifier $\Phi(\cdot)$ (if available) on the base-task only, and then freeze the backbone during novel tasks. For novel classes, the classifier is extended by mean features (*i.e.* prototypes) extracted by the frozen backbone. We follow and employ the scheme as our baseline.

Main idea of our method. Given that the backbone is initially trained for the base task and remains fixed during novel tasks, the main challenges lie in the potential overfitting on the base task and the poor feature extraction ability of the novel classes. We conduct a toy experiment to show these problems and show that the intermediate feature can convey valuable information to tackle these problems. Results are shown in Fig. 3, where we select 4 base classes and 5 novel classes and utilize the t-SNE [37] to visualize the features from different layers. It can be concluded that features extracted by the base-classes trained backbone are overfitted to the base classes and cannot be generalized well on novel classes, *i.e.* the final features are well aggregated, but it is hard to distinguish among unseen classes. (see L_{12} features on the left and the corresponding performance on the right). Compared with the final features (*i.e.* L_{12}), the shallower intermediate features (L_9, L_{10}, L_{11}) are more dispersed, which makes them more discriminative for unseen classes. This is cross-verified by the performance shown on the right side of Fig. 3. We can see that using intermediate features can largely boost the novel classes' performance (up to 15% improvement) with a price of performance decline of the base classes. Based on this observation, we propose to rectify the final feature based on the intermediate feature to better handle both base and novel classes.

4.2 Feature Rectification

We propose a Feature Rectification (FR) module to rectify the final feature based on the relation information from the intermediate features. Specifically, given a pair of final feature X_{N_L} and the intermediate feature X_l, we design an FR module conditioned on the intermediate features:

$$X_{FR} = M_{mix}(\mathrm{cat}(M_F(X_{N_L}), M_I(X_l)), \tag{7}$$

where 'cat' means the concatenate operation. M_{mix}, M_F and M_I are 2-layer MLPs. By introducing the intermediate features, the FR module has access to a reference for extracting the valuable structural information lies in the intermediate features. A naive way to mine such information is to directly pull rectified features closer to shallow features on an instance-by-instance basis. Although such an instance-wise loss can bridge shallow features and rectified features, directly pulling the rectified feature to the shallow one could lead to noisy features since the shallow features are more scattered, as shown in Sec. 4.1.

To effectively transfer discriminative cues among samples from shallow features to rectified features and avoid introducing noise, we propose two relation-based transfer losses at different levels. Firstly, we offer a **Instance-level Relation transfer** (IR) to mine the relation between feature instances. Given a pair of final features $X^1_{N_L}$ and $X^2_{N_L}$, and another pair of intermediate features X^1_l and x^2_l, the IR loss is defined as:

$$\mathcal{L}_{IR} = \delta(\epsilon(X^1_{FR}, X^2_{FR}), \epsilon(X^1_l, x^2_l)), \tag{8}$$

where $\epsilon(\cdot)$ is the Euclidean distance and the $\delta(\cdot)$ denotes the smooth L1 loss. Besides the relation between feature instances, another level of the relation information is the relation between features and class centers. **Class-center Relation Transfer** (CR) is proposed to mine this class-center-level information. Assume at task T_i with $|\mathcal{Y}_i|$ classes and the feature dimension is d. Given a set of class prototypes $P \in \mathbb{R}^{d\times|\mathcal{Y}_i|}$ from the classifier $\Phi(\cdot)$ as well as a pair of final feature $X_{N_L} \in \mathbb{R}^{d\times 1}$ and the intermediate feature X_l ($l \in [1, N_L)$), the corresponding CR loss is formulated as:

$$\mathcal{L}_{CR} = \mathrm{KL}((\sigma(P^\top \overline{X_{FR}}), \sigma(P^\top \overline{X_l})), \tag{9}$$

where $\sigma(\cdot)$ is the softmax function, and $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence. $\overline{X}$ denotes the L2 normalization results of X.

In the above equations, the two losses $\mathcal{L}_{CR}$ and $\mathcal{L}_{IR}$ work on different levels and serve different roles. $\mathcal{L}_{CR}$ works on the relations between a sample and all class centers, focusing on the coarse-grained class-specific patterns that distinguish different classes. Differently, $\mathcal{L}_{IR}$ provides knowledge about fine-grained discriminative cues that contribute to pairwise relations. The proposed relation-based transfer losses are related to previous knowledge distillation works [5, 40, 41]. However, these works do not share the same objective as ours. Our goal is to extract precious transferrable information hidden in various shallow layers hierarchically to rectify the final features while prior distillation methods focus on matching final outputs to teachers. Furthermore, there exists a branch of works using intermediate features [36,62] directly for dense prediction. Differently, we are the first to extract intermediate knowledge to rectify the final feature for a classification task.

Besides, we constrain the discriminative ability of rectified features on base classes to avoid negative rectification. Since the final feature of layer L_{N_L} is strong enough for the base classes, we apply a cosine constraint as:

$$\mathcal{L}_{cos} = \cos(X_{FR}, X_{N_L}), \tag{10}$$

where $\cos(\cdot,\cdot)$ is cosine similarity.

Novel class training. To ensure the rectified features are centered around the class mean prototypes, we further introduce a classification loss for novel classes:

$$\mathcal{L}_{NovCE} = -\mathbb{1}(y \in \cup_{n>1}\mathcal{Y}_n) \log \sigma(P^\top \overline{X_{FR}})^{(y)}, \tag{11}$$

where y is the corresponding ground true label of X_{FR}, $(P^\top \overline{X_{FR}})^{(y)}$ is the y-th element of $(P^\top \overline{X_{FR}})$.

Anti-forgetting training. Another challenge of FSCIL is the forgetting problem. Thanks to our feature rectification design, the only trainable module is the FR module. As the FR module takes the final and the intermediate feature as the input, these kinds of features can be easily modeled by Gaussian distributions [51,71]. Hence, we construct a Gaussian distribution for each class on each layer. The Gaussian distribution of class i features at layer L_l can be denoted as $\mathcal{N}_i^l(\mu_i^l, \Sigma_i^l)$, where Σ_i^l is obtained by calculating the covariance of each feature dimension if $i \in \mathcal{Y}_1$ (base classes). For novel classes, the covariance matrix is computed by averaging the covariance matrices of the top-k most similar base classes. During the incremental learning phases, we utilize these Gaussian distributions to sample features, which serve as inputs from old tasks to the FR module, assisting the FR module in retaining previously learned knowledge.

Overall training objectives. The total loss function for our feature rectification framework is as follows:

$$\mathcal{L} = \beta_{cos}(\mathcal{L}_{cos} + \mathcal{L}_{NovCe}) + \beta_{CR}\mathcal{L}_{CR} + \beta_{IR}\mathcal{L}_{IR}, \tag{12}$$

where β_{cos}, β_{CR} and β_{IR} are trade-off parameters.

Multi-layer knowledge ensemble. Since there exists hierarchical knowledge in different layers, we mine such information by knowledge ensemble from different layers. Specifically, we incorporate an FR module with each selected layer, resulting in a total of n_{FR} FR modules. During the training phase, these FR modules are optimized simultaneously. When evaluation, FR modules at layer l are employed to produce multiple rectified features denoted as X_{FR}^l. These rectified features are subsequently input into the shared classifier P for generating multiple predictions: $\text{Pred}_l = \sigma(P^\top \overline{X_{FR}^l})$. The final prediction is computed as the average of each FR-branch: $\text{Pred} = \frac{1}{n_{FR}} \sum_{l=1}^{n_{FR}} \text{Pred}_l$.

5 Experiments

5.1 Experiment Setup

Datasets. Following previous works, we conduct experiments on three popular datasets of FSCIL: CIFAR-100 [28], *mini*ImageNet [46] and CUB-200 [55]. For statistics of the three datasets, please refer to the supplementary material.

Table 1: Comparison with FSCIL methods on *mini*ImageNet. The best and second best results are marked **bolded** and underlined. Methods with†are reported with the original results in their paper. Other results are reimplemented by us.

Methods	**Num. of Params.(M)**	*aAcc* **in each task (%)**										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	*aAcc*	*gAcc*
TOPIC [52]†	-	61.3	50.1	45.2	41.2	37.5	35.5	32.2	29.5	24.4	39.64	-
IDLVQ [6]†	-	64.8	59.8	55.9	52.6	49.9	47.6	44.8	43.1	41.8	51.16	-
MetaFSCIL [8]†	-	72.0	68.0	63.8	60.3	57.6	55.2	53.0	50.8	49.2	58.85	-
DF-Replay [35]†	-	71.8	67.1	63.2	59.8	57.0	54.0	51.6	49.5	48.2	58.02	-
FCIL [18]†	-	76.34	71.4	67.1	64.1	61.3	58.5	55.7	54.1	52.8	62.37	-
Self-promoted [73]	12.34	63.6	64.1	59.5	55.5	52.1	49.2	45.6	43.1	40.9	52.62	41.53
CEC [63]	12.12	72.1	67.0	63.0	59.6	56.9	53.9	51.4	49.5	47.8	57.91	48.64
LIMIT [70]	12.29	73.3	67.6	63.2	59.5	56.7	53.7	51.4	49.7	48.2	58.15	49.47
FACT [69]	11.26	76.3	71.3	66.9	62.9	59.6	56.3	53.5	51.1	49.2	60.78	49.56
CLOM [76]	14.22	72.1	66.7	62.7	59.2	56.1	53.0	50.2	48.3	46.7	57.23	46.11
C-FSCIL [20]	12.79	70.6	63.8	59.3	56.4	54.3	51.8	48.9	47.3	46.1	55.40	50.46
BiDistFSCIL [66]	11.50	74.6	70.4	66.6	62.8	60.5	57.2	54.7	53.0	51.9	61.32	53.49
S3C [24]	0.34	77.0	72.1	68.2	65.1	62.0	58.8	56.0	53.6	51.9	62.74	53.70
Regularizer [3]	26.30	81.9	72.7	68.7	64.1	57.4	53.5	53.8	52.3	43.1	62.82	55.62
SubNetwork [25]	22.38	78.1	73.5	69.8	66.2	63.2	60.3	57.6	55.7	54.5	64.31	55.88
SAVC [49]	12.12	81.0	76.2	71.9	71.4	65.6	62.1	59.2	57.0	55.5	66.66	56.35
ALICE [42]	11.20	81.1	71.1	67.8	64.1	62.4	60.1	57.9	56.8	56.0	64.14	59.72
NC-FSCIL [59]	15.87	84.2	76.1	72.1	68.1	67.3	63.9	61.9	59.6	57.9	67.90	60.93
YourSelf (Ours)	11.57	84.0	77.6	73.7	70.0	68.0	64.9	62.1	59.8	**59.0**	**68.80**	**62.20**

Training and testing protocols. For training protocols, we follow prior works [3, 20, 24, 25, 42, 49, 59, 66] to split a large proportion of the training set to serve as the base task, and split the rest into small incremental tasks. Specifically, for CIFAR-100 and *mini*ImageNet, the first task contains 60 classes out of the total 100 classes, and the rest 40 classes are split into 8 tasks, following a 5-way 5-shot scheme. For CUB-200, there are 100 classes in the first task and the number of novel tasks is 10 (in a 10-way 5-shot setting). As for the testing, we follow prior works to report the *average accuracy (aAcc)*. In addition, we also provide detailed discussions on our metric *generalized accuracy (gAcc)*. We further provide detailed results of novel classes only in our supplementary material.

Implementation details. For all experiments below, we use the ViT suggested by Touvron et, al. [53]. We adjust stuff like the dimensions of the backbone to match the number of parameters (around 10M) as the ResNet18 [19] which is widely used in this topic. The Feature Rectification (FR) is a lightweight module and uses popular MLPs implementation [53] with only 0.6M parameters. To ensure a fair comparison, we also re-implement other methods with the same backbone as ours, but found the performance is rarely improved. Please refer to Supplementary material for more details. We follow previous works to **train from scratch** on CIFAR-100 and *mini*ImageNet dataset. Prior works [6, 8, 35, 52, 63, 73] adopt an ImageNet-pretrained ResNet for the CUB-200 dataset. To make a fair comparison, we also pretrained our backbone on the ImageNet [12] to evaluate on CUB-200. For evaluation, we reproduced 14 methods based on the public codes and conducted in-depth comparisons in the following sections.

5.2 Comparison with State-of-the-art methods

Experimental results on *mini*ImageNet are shown in Table 1. We report the classical accuracy ($aAcc$) in each section as well as its averaged results and our proposed $gAcc$. From the table, we can conclude that our proposed method

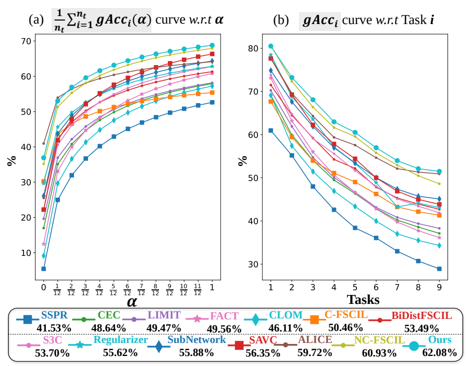

Fig. 5: FSCIL results on *mini*ImageNet. (a): The $gAcc$ curve (averaged across all n_t tasks) $vs.$ α. (b): The $gAcc$ AUC of each task $\mathcal{T}_i$ (Eq. (5))$vs.$ i. In the legend, we show the AUC value (Eq. (6)) of each method. Best viewed in color.

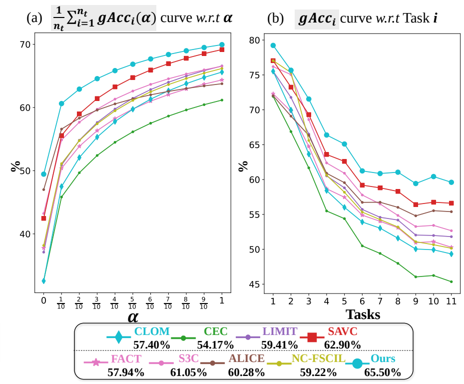

Fig. 6: FSCIL results on CUB-200. (a): The $gAcc$ curve (averaged across all n_t tasks) $vs.$ α. (b): The $gAcc$ AUC of each task $\mathcal{T}_i$ (Eq. (5))$vs.$ i. In the legend, we show the AUC value (Eq. (6)) of each method. Best viewed in color.

outperforms other methods by a large margin. Specifically, under the classical evaluation metric $aAcc$, the proposed method achieves 68.80% accuracy, which is 0.90%, 4.66%, 2.14% higher than SOTA methods NC-FSCIL [59], ALICE [42] and SAVC [49]. The improvement is non-trivial considering that *mini*ImageNet is a complex dataset.

From a more balanced perspective, we also report $gAcc$ on the Table 1 and Fig. 5. In the left figure of Fig. 5, except for the method ALICE when $\alpha = 0$, our method outperforms most other methods by a clear margin(at leat 1.27%) and achieves the highest AUC of all methods. This indicates that our method achieves a more balanced base-novel performance. To see raw numbers of the accuracy of novel classes, please refer to the supplementary material.

Besides, Fig. 5 shows that methods that perform well at $\alpha = 1$ do not necessarily perform well when α is smaller. For example, although CLOM achieves 57.23% average accuracy ($\alpha = 1$) and is 1.83% higher than method C-FSCIL (55.4%), we can see the performance drops significantly when α is small. It is mainly because CLOM focuses on the base-classes accuracy and performs poorly on novel classes. We can observe that the AUC of the CLOM is much smaller (4.35%) than that of C-FSCIL. The results confirm the necessity of our proposed $gAcc$. Similar phenomena can be found when comparing other methods like SAVC and ALICE, and we can obtain similar conclusions on the other two datasets including CUB-200 and CIFAR-100.

Results on CUB-200 dataset are shown in Table 2 and Fig. 6. We can see from Table 2 that our method is better than any other existing methods in terms of $aAcc$ and $gAcc$ (1.52%@$aAcc$ and 2.6%@$gAcc$ improvement over the strong method SAVC). Results on CIFAR-100 dataset are shown in Table 3. For the traditional metric $aAcc$, our method achieves comparable performance with SOTA methods, while our method outperforms all other methods by a clear margin when considering the newly introduced metric $gAcc$. Please refer to Sec A5 of the supplementary material for more experiment results.

Table 2: Comparison with FSCIL methods on CUB-200. The best and second best results are marked **bolded** and <u>underlined</u>. Methods with † are reported with the original results in their paper. Other results are reimplemented by us.

Methods	Num. of Params.(M)	*aAcc* in each task (%) 1	2	3	4	5	6	7	8	9	10	11	*aAcc*	*gAcc*
TOPIC [52]†	-	68.7	42.5	54.8	50.0	45.3	41.4	38.4	35.4	32.2	28.3	26.3	43.92	-
DF-Replay [35]†	-	75.9	72.1	68.6	63.8	62.6	59.1	57.8	55.9	55.0	53.6	52.4	61.52	-
BiDistFSCIL [66]†	-	79.1	75.4	72.8	69.1	67.5	65.1	64.0	63.5	61.9	61.5	60.9	67.34	-
MetaFSCIL [8]†	-	75.9	72.4	68.8	64.8	63.0	60.0	58.3	56.8	54.8	53.8	52.6	61.93	-
SubNetwork [25]†	-	78.7	74.6	71.4	67.5	65.4	62.6	61.1	59.4	57.5	57.2	56.7	64.68	-
DSN [58]†	-	76.1	72.2	69.6	66.7	64.4	62.1	60.2	58.9	57.0	55.1	54.2	63.31	-
Self-promoted [73]	12.36	62.6	54.7	51.6	47.3	44.3	40.9	38.8	39.0	35.9	35.1	33.5	44.07	40.88
CEC [63]	12.34	75.7	71.4	68.1	64.0	62.0	58.8	57.3	55.5	53.6	53.2	52.1	61.17	54.16
LIMIT [70]	12.33	79.5	76.4	73.2	69.2	67.1	64.3	62.8	61.9	59.8	59.1	57.6	66.58	59.40
CLOM [76]	18.89	79.5	75.9	72.0	68.2	65.7	63.4	62.0	60.3	58.5	57.9	56.9	65.61	57.40
FACT [69]	11.34	76.1	73.4	70.4	66.5	64.8	62.5	61.1	59.9	57.9	57.5	56.5	50.29	57.93
NC-FSCIL [59]	15.91	81.1	76.4	72.8	69.3	66.1	63.7	62.2	61.0	59.0	58.0	57.3	66.17	59.22
ALICE [42]	41.71	75.8	70.1	68.4	65.2	63.7	61.2	60.7	59.7	58.5	58.7	58.4	63.77	60.28
S3C [24]	13.06	80.2	76.5	72.9	69.1	66.9	64.3	62.7	61.0	59.5	59.2	58.3	66.53	61.05
SAVC [49]	24.29	81.1	77.6	75.0	71.2	69.6	66.7	65.8	64.8	62.9	62.7	<u>62.2</u>	<u>68.33</u>	<u>62.90</u>
YourSelf (Ours)	11.90	83.4	77.0	75.3	72.2	69.0	66.8	66.0	65.6	64.1	64.5	**63.6**	**69.85**	**65.50**

Table 3: Comparison results on CIFAR-100. The best and second best results are marked **bolded** and <u>underlined</u>. All results of other methods are reimplemented by us. See more results in the supplementary material.

Methods	Num. of Params.(M)	*aAcc* in each task (%) 1	2	3	4	5	6	7	8	9	*aAcc*	*gAcc*
CEC [63]	0.296	73.2	69.1	65.4	61.2	58.0	55.5	53.2	51.3	49.2	59.56	50.85
LIMIT [70]	0.296	76.1	71.7	67.3	63.2	60.1	57.3	55.2	53.0	50.8	61.64	52.16
CLOM [76]	0.355	74.2	69.8	66.1	62.2	59.0	56.1	54.0	51.9	49.8	60.35	50.36
C-FSCIL [20]	12.79	76.8	72.0	67.2	63.1	59.7	56.7	54.4	51.9	49.7	61.28	50.62
SAVC [49]	0.842	78.4	72.7	68.4	64.0	61.2	58.3	56.0	54.1	51.7	62.76	53.29
FACT [69]	0.282	78.5	72.7	68.8	64.5	61.3	58.6	56.8	54.4	52.3	63.12	54.56
SubNet [25]	22.39	80.0	75.7	71.6	67.8	64.7	61.4	59.3	57.1	55.1	64.85	56.63
S3C [24]	0.341	78.0	73.9	70.2	66.1	63.3	60.1	58.3	56.7	54.0	64.51	58.04
ALICE [42]	45.28	80.3	71.9	67.0	63.2	60.7	58.3	57.3	55.5	53.7	63.10	59.72
NC-FSCIL [59]	15.93	82.9	77.5	73.7	69.0	65.6	62.0	59.7	57.8	<u>55.4</u>	**67.06**	<u>59.73</u>
Ours	11.68	82.9	76.3	72.9	67.8	65.2	62.0	60.7	58.8	**56.6**	<u>67.02</u>	**60.32**

5.3 Ablation studies

Effectiveness of different losses. Ablation experiments of the effectiveness of $\mathcal{L}_{IR}$ and $\mathcal{L}_{CR}$ are conducted among CIFAR-100, *mini*ImageNet, and CUB-200 datasets. Experimental results are shown in Table 4 and Table 5. The detailed *gAcc* *vs*. α as shown in Fig. 7. Results on CIFAR-100 show that introducing the FR module can effectively enhance the *gAcc* by 3.28%. After further introducing the class-relation (CR) transfer loss, the performance at *gAcc* is further improved by 1.07%. Adding the instance-relation (IR) transfer loss will further boost the *gAcc* by 1.91%. The effectiveness of the proposed losses is also verified on the *mini*ImageNet dataset. For *mini*ImageNet dataset, the introduction of FR modules brings 3.02% improvement at *gAcc*. Applying loss $\mathcal{L}_{CR}$ and $\mathcal{L}_{IR}$ can improve *gAcc* by 0.49% and 0.82%, respectively. It is worth mentioning that $\mathcal{L}_{CR}$ and $\mathcal{L}_{IR}$ can also boost the *aAcc*. Specifically, on the *mini*ImageNet, adding these two losses results in an increase at *aAcc* performance of 1.96%. On CIFAR-100, the *aAcc* is improved by about 0.76% with the help of these two losses. See more experiments in the supplementary material.

Effectiveness of multi-layer knowledge ensemble. In Sec. 4.2, we propose to exploit hierarchical knowledge from multiple layers for effective feature rectification. In practice, we select layers 8, 9 and 10 for the ensemble. Here, we take a more detailed discussion on each layer before the ensemble of multiple

Table 4: Ablation studies on CIFAR-100.

FR	$\mathcal{L}_{Cos}+\mathcal{L}_{NovCE}$	$\mathcal{L}_{CR}$	$\mathcal{L}_{IR}$	aAcc. in each task (%) 1	2	3	4	5	6	7	8	9	$aAcc$	$gAcc$
				82.9	77.8	72.9	68.4	64.6	61.4	58.7	56.4	54.1	66.35	54.20
✓	✓			82.9	76.7	73.1	67.9	64.8	60.8	59.2	57.2	53.9	66.26	57.48
✓	✓	✓		82.9	76.1	72.4	67.8	64.5	61.3	59.6	57.3	54.9	66.33	58.55
✓	✓	✓	✓	82.9	75.7	72.5	67.4	64.8	61.9	60.4	58.7	56.3	67.02	60.32

Table 5: Ablation studies on *mini*ImageNet.

FR	$\mathcal{L}_{Cos}+\mathcal{L}_{NovCE}$	$\mathcal{L}_{CR}$	$\mathcal{L}_{IR}$	aAcc. in each task (%) 1	2	3	4	5	6	7	8	9	$aAcc$	$gAcc$
				84.0	79.4	75.0	71.1	67.8	64.4	61.3	58.8	56.9	68.74	57.87
✓	✓			84.0	74.0	69.7	66.8	66.1	63.5	60.8	58.6	58.1	66.84	60.89
✓	✓	✓		84.0	77.7	73.6	69.9	67.4	64.6	61.6	59.8	58.7	68.60	61.38
✓	✓	✓	✓	84.0	77.6	73.7	70.0	68.0	64.9	62.1	59.8	59.0	68.80	62.20

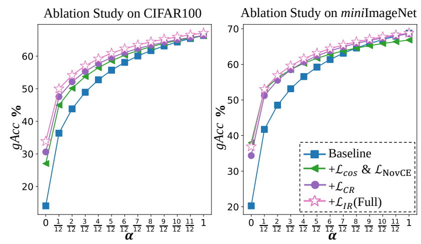

Fig. 7: Ablation studies on two datasets, evaluating using $gAcc$.

layers. We conducted experiments on the *mini*ImageNet, and the results are shown in Table 6. It is clear that for each layer, the corresponding FR module effectively calibrates the features and achieves improvement from 2.23% to 3.54% at $gAcc$. Furthermore, the introduction of the FR module does not harm the $aAcc$. Considering the fact that different layers contain hierarchical knowledge, the multi-layer ensemble is proposed. The results in the table show that the ensemble is effective in improving the $gAcc$ while maintaining $aAcc$.

Table 6: Detailed performance of each layer. Results are from *mini*Imagenet dataset. 'l-Raw': vanilla ViT features from layer L_l. 'l-Recti' : Rectified features using the proposed FR module. 'Ensemble' : Ensembled results from 'l-Recti', $l \in \{8, 9, 10\}$.

Layers	$aAcc$ in each session (%) 0	1	2	3	4	5	6	7	8	$aAcc$	$gAcc$
8-Raw	78.9	73.2	69.1	65.4	62.9	60.0	57.3	55.2	54.0	64.01	56.61
8-Recti	80.7	73.4	70.1	66.6	64.4	61.5	59.3	57.0	56.5	65.52	60.15
9-raw	82.5	77.3	73.2	69.7	65.3	63.7	60.9	58.9	57.5	67.66	59.33
9-Recti	83.0	76.6	73.2	69.7	67.5	64.6	61.9	59.9	59.0	68.39	61.90
10-raw	83.8	78.6	74.4	70.7	66.4	64.3	61.5	59.3	57.7	68.55	58.90
10-Recti	84.3	78.2	74.1	70.4	68.2	64.8	61.9	59.5	58.4	68.84	61.13
Ensemble	84.0	77.6	73.7	70.0	68.0	64.9	62.1	59.8	59.0	68.80	62.20

6 Conclusion

In this work, we have revisited the challenges of the FSCIL, showing the limitation of the existing metric and introducing a novel metric denoted as *generalized accuracy* ($gAcc$) for the comprehensive evaluation of FSCIL methods from a more balanced perspective. Furthermore, we incorporate the ViT into the FSCIL framework and unlock the potential of the ViT intermediate layer features. To achieve this, we have designed a feature rectification (FR) module aimed at integrating the valuable information from the intermediate layer features into the rectified feature space. Using the new metric $gAcc$, we analyze the existing open-sourced FSCIL methods from a more balanced perspective and show the superiority of our method on three popular datasets.

Acknowledgements

This work was supported partially by the National Key Research and Development Program of China (2023YFA1008503), NSFC(U21A20471, 62206315), Guangdong NSF Project (No. 2023B1515040025, 2020B1515120085, 2024A15150-10101), Guangzhou Basic and Applied Basic Research Scheme(2024A04J4067).

References

1. Agarwal, A., Banerjee, B., Cuzzolin, F., Chaudhuri, S.: Semantics-driven generative replay for few-shot class incremental learning. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
2. Ahmad, T., Dhamija, A.R., Cruz, S., Rabinowitz, R., Li, C., Jafarzadeh, M., Boult, T.E.: Few-shot class incremental learning leveraging self-supervised features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
3. Akyürek, A.F., Akyürek, E., Wijaya, D.T., Andreas, J.: Subspace regularizers for few-shot class incremental learning. In: International Conference on Learning Representations (2022)
4. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: Proceedings of the European Conference on Computer Vision (2018)
5. Chen, H., Wang, Y., Xu, C., Xu, C., Tao, D.: Learning student networks via feature embedding. IEEE Transactions on Neural Networks and Learning Systems (2020)
6. Chen, K., Lee, C.G.: Incremental few-shot learning via vector quantization in deep embedded space. In: International Conference on Learning Representations (2020)
7. Cheraghian, A., Rahman, S., Fang, P., Roy, S.K., Petersson, L., Harandi, M.: Semantic-aware knowledge distillation for few-shot class-incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
8. Chi, Z., Gu, L., Liu, H., Wang, Y., Yu, Y., Tang, J.: Metafscil: A meta-learning approach for few-shot class incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022)
9. 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)
10. Cui, Y., Deng, W., Xu, X., Liu, Z., Liu, Z., Pietikäinen, M., Liu, L.: Uncertainty-guided semi-supervised few-shot class-incremental learning with knowledge distillation. IEEE Transactions on Multimedia (2022)
11. Cui, Y., Xiong, W., Tavakolian, M., Liu, L.: Semi-supervised few-shot class-incremental learning. In: IEEE International Conference on Image Processing (ICIP) (2021)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2009)
13. Deng, Y., Xiang, X.: Expanding hyperspherical space for few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)

14. 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 (2021)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
16. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: Proceedings of the European Conference on Computer Vision (2020)
17. Fu, S., Yan, J., Gao, Y., Xie, X., Zheng, W.S.: Asag: Building strong one-decoder-layer sparse detectors via adaptive sparse anchor generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6328–6338 (2023)
18. Gu, Z., Xu, C., Yang, J., Cui, Z.: Few-shot continual infomax learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
19. 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 (2016)
20. Hersche, M., Karunaratne, G., Cherubini, G., Benini, L., Sebastian, A., Rahimi, A.: Constrained few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
21. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
22. Ji, Z., Hou, Z., Liu, X., Pang, Y., Li, X.: Memorizing complementation network for few-shot class-incremental learning. IEEE Transactions on Image Processing (2023)
23. Jiang, W., Li, D., Hu, M., Zhai, G., Yang, X., Zhang, X.P.: Few-shot class-incremental learning with prior knowledge. arXiv preprint arXiv:2402.01201 (2024)
24. Kalla, J., Biswas, S.: S3c: Self-supervised stochastic classifiers for few-shot class-incremental learning. In: European Conference on Computer Vision (2022)
25. Kang, H., Yoon, J., Madjid, S.R.H., Hwang, S.J., Yoo, C.D.: On the soft-subnetwork for few-shot class incremental learning. In: International Conference on Learning Representations (2023)
26. Kim, D.Y., Han, D.J., Seo, J., Moon, J.: Warping the space: Weight space rotation for class-incremental few-shot learning. In: International Conference on Learning Representations (2023)
27. Kim, S., Jeong, Y., Park, J.S., Yoon, S.W.: Mics: Midpoint interpolation to learn compact and separated representations for few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
28. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
29. Lee, K., Lee, K., Shin, J., Lee, H.: Overcoming catastrophic forgetting with unlabeled data in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
30. Li, J., Dong, S., Gong, Y., He, Y., Wei, X.: Analogical learning-based few-shot class-incremental learning. IEEE Transactions on Circuits and Systems for Video Technology (2024)

31. Li, Y.M., Huang, W.J., Wang, A.L., Zeng, L.A., Meng, J.K., Zheng, W.S.: Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. arXiv preprint arXiv:2406.08877 (2024)
32. Li, Y.M., Zeng, L.A., Meng, J.K., Zheng, W.S.: Continual action assessment via task-consistent score-discriminative feature distribution modeling. IEEE Transactions on Circuits and Systems for Video Technology (2024)
33. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence (2017)
34. Lin, K.Y., Ding, H., Zhou, J., Peng, Y.X., Zhao, Z., Loy, C.C., Zheng, W.S.: Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. arXiv preprint arXiv:2403.01560 (2024)
35. Liu, H., Gu, L., Chi, Z., Wang, Y., Yu, Y., Chen, J., Tang, J.: Few-shot class-incremental learning via entropy-regularized data-free replay. In: European Conference on Computer Vision (2022)
36. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
37. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
38. Mazumder, P., Singh, P., Rai, P.: Few-shot lifelong learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)
39. Pan, Z., Yu, X., Zhang, M., Gao, Y.: Ssfe-net: Self-supervised feature enhancement for ultra-fine-grained few-shot class incremental learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2023)
40. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
41. Passalis, N., Tzelepi, M., Tefas, A.: Heterogeneous knowledge distillation using information flow modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
42. 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 (2022)
43. Peng, Y.X., Jiao, J., Feng, X., Zheng, W.S.: Consistent discrepancy learning for intra-camera supervised person re-identification. IEEE Transactions on Multimedia **25**, 2393–2403 (2022)
44. 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)
45. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
46. 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 (2015)
47. Shi, G., Chen, J., Zhang, W., Zhan, L.M., Wu, X.M.: Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In: Advances in neural information processing systems (2021)

48. Smith, J., Hsu, Y.C., Balloch, J., Shen, Y., Jin, H., Kira, Z.: Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
49. 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 (2023)
50. Tang, Y.M., Peng, Y.X., Zheng, W.S.: Learning to imagine: Diversify memory for incremental learning using unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
51. Tang, Y.M., Peng, Y.X., Zheng, W.S.: When prompt-based incremental learning does not meet strong pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
52. 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 (2020)
53. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning (2021)
54. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need? In: International Conference on Learning Representations (2021)
55. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
56. Wang, Q.W., Zhou, D.W., Zhang, Y.K., Zhan, D.C., Ye, H.J.: Few-shot class-incremental learning via training-free prototype calibration. In: Advances in Neural Information Processing Systems (2023)
57. Yan, S., Xie, J., He, X.: Der: Dynamically expandable representation for class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
58. Yang, B., Lin, M., Zhang, Y., Liu, B., Liang, X., Ji, R., Ye, Q.: Dynamic support network for few-shot class incremental learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
59. 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: International Conference on Learning Representations (2023)
60. Yao, G., Zhu, J., Zhou, W., Li, J.: Few-shot class-incremental learning based on representation enhancement. Journal of Electronic Imaging (2022)
61. Yu, L., Twardowski, B., Liu, X., Herranz, L., Wang, K., Cheng, Y., Jui, S., Weijer, J.v.d.: Semantic drift compensation for class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
62. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution vision transformer for dense predict. Advances in Neural Information Processing Systems (2021)
63. 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 (2021)
64. Zhang, J., Zhang, J., Ghosh, S., Li, D., Tasci, S., Heck, L., Zhang, H., Kuo, C.C.J.: Class-incremental learning via deep model consolidation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (2020)

65. Zhao, B., Xiao, X., Gan, G., Zhang, B., Xia, S.T.: Maintaining discrimination and fairness in class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
66. 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 (2023)
67. Zheng, G., Zhang, A.: Few-shot class-incremental learning with meta-learned class structures. In: International Conference on Data Mining Workshops (ICDMW) (2021)
68. Zheng, W.S., Yan, J., Peng, Y.X.: A versatile framework for multi-scene person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
69. 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 (2022)
70. Zhou, D.W., Ye, H.J., Ma, L., Xie, D., Pu, S., Zhan, D.C.: Few-shot class-incremental learning by sampling multi-phase tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
71. Zhu, F., Zhang, X.Y., Wang, C., Yin, F., Liu, C.L.: Prototype augmentation and self-supervision for incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
72. Zhu, J., Yao, G., Zhou, W., Zhang, G., Ping, W., Zhang, W.: Feature distribution distillation-based few shot class incremental learning. In: the 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI) (2022)
73. Zhu, K., Cao, Y., Zhai, W., Cheng, J., Zha, Z.J.: Self-promoted prototype refinement for few-shot class-incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
74. Zhu, K., Zhai, W., Cao, Y., Luo, J., Zha, Z.J.: Self-sustaining representation expansion for non-exemplar class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
75. Zhu, M., Zhu, Z., Chen, S., Chen, C., Wu, B.: Enhanced few-shot class-incremental learning via ensemble models. arXiv preprint arXiv:2401.07208 (2024)
76. Zou, Y., Zhang, S., Li, Y., Li, R.: Margin-based few-shot class-incremental learning with class-level overfitting mitigation. In: Advances in neural information processing systems (2022)