Few-shot Class-incremental Learning for Retinal Disease Recognition

Jinghua Zhang, Peng Zhao, Yongkun Zhao, Chen Li, and Dewen Hu

***Abstract*— *Few-Shot Class-Incremental Learning* (FSCIL) techniques are essential for developing *Deep Learning* (DL) models that can continuously learn new classes with limited samples while retaining existing knowledge. This capability is particularly crucial for the DL-based retinal disease diagnosis system, where acquiring large annotated datasets is challenging, and disease phenotypes evolve over time. This paper introduces Re-FSCIL, a novel framework for *Few-Shot Class-Incremental Retinal Disease Recognition* (FSCIRDR). Re-FSCIL integrates the RETFound model with a fine-grained module, employing a forward-compatible training strategy to improve adaptability, supervised contrastive learning to enhance feature discrimination, and feature fusion for robust representation quality. We convert existing datasets into the FSCIL format and reproduce numerous representative FSCIL methods to create two new benchmarks, RFMiD38 and JSIEC39, specifically for FSCIRDR. Our experimental results demonstrate that Re-FSCIL achieves *State-of-the-art* (SOTA) performance, significantly surpassing existing FSCIL methods on these benchmarks.**

***Index Terms*— Retinal disease, Class-incremental learning, Few-shot learning, Deep learning, Foundation model**

I. Introduction

According to the World Health Organization's World Report on Vision 2019, approximately 2.2 billion people globally are affected by visual impairment, with at least 1 billion of these cases being preventable or yet to be addressed [1]. Retinal diseases contribute to these impairments, highlighting the critical need for early screening and diagnosis [2]. The retina at the back of the eye plays a vital role in vision by converting incoming light into electrical signals and transmitting them to the brain via the optic nerve [3]. Due to its unique characteristics, the retina can reveal not only eye-specific diseases but also broader physiological conditions, particularly those related to the circulatory system and neurological disorders [4]. Color fundus photography has achieved significant success in diagnosing various chronic systemic diseases such as diabetes, hypertension, and other cardiovascular conditions [5], [6]. Given its ability to non-invasively observe retinal microcirculation, it is also used for the identification of retinal diseases.

However, current diagnostic methods primarily rely on doctors' experience and subjective judgment, leading to inconsistent results [7]. Doctors analyzing retinal images may produce varying diagnoses due to differences in experience, fatigue, or other subjective factors, adding uncertainty and potential risk to medical decisions. Moreover, in underdeveloped regions, the scarcity of medical resources, the variety of diseases, and the demand for large-scale screening further complicate diagnosis [8], [9]. These areas often lack adequate medical equipment and trained professionals to handle the high screening workload, resulting in many patients not receiving timely and accurate diagnosis and treatment. This situation highlights deficiencies in the healthcare system regarding technology, resources, and training, calling for increased support and improvements to enhance diagnostic consistency and coverage. To address these limitations, the application of computer vision-based artificial intelligence technologies has increased, providing fast, objective, and efficient solutions for the diagnosis of retinal diseases [10]–[12].

Existing vision-based *Computer-assisted Diagnosis* (CAD) for retinal diseases can be broadly divided into two categories: traditional methods [13]–[15] and *Deep Learning* (DL)-based vision methods [16]–[19]. Traditional methods, which address issues in retinal recognition, generally rely on expert domain knowledge to design specialized recognition strategies. However, these strategies are tailored to specific problems, leading to poor generalization performance. In contrast, current DL methods have resolved many shortcomings of traditional methods by improving generalization through automatic feature learning [20]–[24], but they still have some drawbacks. Despite achieving many successes in accuracy and efficiency, DL-based retinal disease diagnosis technologies face numerous challenges in practical applications. One major issue is the reliance on large annotated datasets for training. In the medical imaging field, especially in collecting and annotating retinal images, data collection is difficult and time-consuming, resulting in limited data availability. Additionally, there is a scarcity of expert-labeled data, and training high-quality models requires a substantial amount of annotated data. This

This work was supported by the National Natural Science Foundation of China under Grant 62306323 and the China Scholarship Council under Grant 202206110005.

Jinghua Zhang (zhangjingh@foxmail.com) and Dewen Hu (dwhu@nudt.edu.cn) are with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China, and Jinghua Zhang is also affiliated with CMVS, University of Oulu, Finland. Peng Zhao (zp2570288754@gmail.com) is with the College of Systems Engineering, National University of Defense Technology, Changsha, China. Yongkun Zhao (y.zhao22@imperial.ac.uk) is with the Department of Bioengineering, Imperial College London, UK. Chen Li (lichen@bmie.neu.edu.cn) is with the College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China.

Jinghua Zhang and Peng Zhao contribute equally. Dewen Hu and Chen Li are the corresponding authors.

dependence on large datasets limits the rapid dissemination of the technology and creates a demand for models that can effectively train on limited samples. Moreover, most existing DL methods for retinal diagnosis are static; once trained, they struggle to adapt to new data or changes in the environment. However, medical data and disease phenotypes are constantly evolving, and updates in pathological information and diagnostic standards require diagnostic systems to dynamically adapt to these changes.

In this context, *Few-Shot Class-Incremental Learning* (FSCIL) techniques become particularly important as they enable models to continuously learn new class knowledge and retain existing knowledge even with limited training samples [25]. Therefore, exploring the application of FSCIL in retinal disease diagnosis can significantly reduce data annotation costs and decrease computational demands due to retraining, thereby facilitating the widespread use and development of DL-based retinal disease diagnosis technologies. To understand how *Few-Shot Class-Incremental Retinal Disease Recognition* (FSCIRDR) works, we provide the illustration in Fig. 1.

This paper is dedicated to the study of FSCIRDR. We propose a framework named Re-FSCIL, which integrates the foundational RETFound model with a fine-grained module to enhance retinal disease recognition. Re-FSCIL employs a "Feature Embedding + Nearest Mean Classifier" strategy to mitigate overfitting and catastrophic forgetting while ensuring adaptability to new classes. To enhance the model's adaptability to future incremental classes, we incorporate a forward-compatible training strategy by generating virtual classes. These virtual classes simulate potential future classes, allowing the model to effectively foresee and adapt to new classes. Additionally, we introduce supervised contrastive learning to our framework. This technique aims to minimize intra-class differences and maximize inter-class separation, thereby improving the model's ability to distinguish between different retinal diseases with subtle visual differences. Furthermore, we enhance feature extraction by integrating features from the pre-trained RETFound model. RETFound, pre-trained on a large-scale dataset of retinal images, provides robust and representative features. We ensure comprehensive and accurate feature representation by fusing these features with those extracted from our fine-grained module. Overall, our comprehensive framework addresses the challenges of FSCIRDR by incorporating forward-compatible training, supervised contrastive learning, and feature fusion, providing a scalable and robust solution for continuous learning in retinal disease classification. The key contributions of this work can be summarized as follows:

- **Fine-grained Module for Learning Better Features:** Considering the high similarity among different retinal diseases, we developed a specifically designed fine-grained module to extract superior features, enhancing the model's capability in the recognition process.
- **Novel Method for FSCIRDR:** We introduce the Re-FSCIL framework for FSCIRDR, which integrates the RETFound model with the fine-grained module. Our framework employs a forward-compatible strategy to improve adaptability for future classes and utilizes supervised contrastive learning for better feature discrimination. To our knowledge, this is the first method to explore FSCIL in retinal disease recognition.
- **New Benchmarks for FSCIRDR:** Based on the existing retinal disease datasets, we constructed two datasets (RFMiD38 and JSIEC39) for FSCIL and established testing protocols. We reproduced representative methods on these two datasets, creating new benchmarks for FSCIRDR. This work sets new standards and fosters further exploration within the FSCIRDR community.
- **State-of-the-Art Performance:** We demonstrate that the proposed method significantly surpasses existing advanced FSCIL methods on the RFMiD38 and JSIEC39 datasets, establishing new *State-of-the-art* (SOTA) performance benchmarks.

The structure of this paper is organized as follows: Section II introduces related work. Section III details the problem setting, challenges, and our method. Section IV discusses datasets and implementation details and provides results and analysis of the experiments. Finally, conclusions are presented in Section V.

II. Related Work

This section discusses the relevant knowledge encompassing retinal disease recognition, FSCIL, and foundation models.

A. Retinal Disease Recognition

CAD techniques have been widely implemented for the detection and diagnosis of retinal diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration. Many studies [16]–[18], [26]–[29] have enhanced the classification and diagnostic accuracy of these diseases through the use of DL-based algorithms. For instance, the binary classification methods proposed by *Gulshan et al.* [26], ensemble strategies introduced by *Zhang et al.* [27], and the solution for long-tailed data distribution in retinal disease diagnosis proposed by *Ju et al.* [28] and *Zhou et al.* [29]. Additionally, the ADINet framework has been proposed by *Meng and Shin'ichi* [30], integrating class label prediction and attribute prediction into an incremental learning framework. By applying knowledge distillation and attribute distillation techniques, ADINet effectively enhances performance. Although incremental technologies have been applied to diagnose retinal diseases, the application of FSCIL in this field has not been studied. FSCIL is particularly crucial for retinal disease diagnosis as it effectively handles newly emerging classes with scarce samples. This technique allows diagnostic systems to retain knowledge of existing classes while learning new ones, facilitating adaptation to changes in the medical environment. It enhances the diagnostic efficiency for rare diseases and reduces dependence on large-scale annotated data, thus improving the system's flexibility and adaptability in practical applications.

B. Few-shot Class-incremental Learning

FSCIL is a new topic in machine learning aimed at designing algorithms that can continuously learn new class

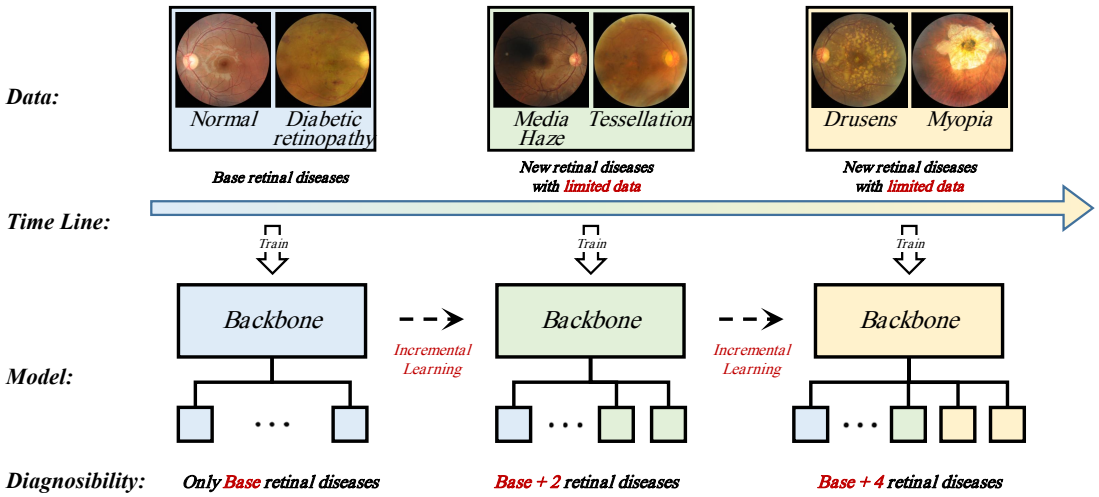


Fig. 1: Illustration of the FSCIRDR process. As time progresses, retinal disease data continuously updates and evolves. Given the high costs associated with data collection and annotation, it is essential for the model to learn new classes with only a few training examples continuously.

knowledge from a limited number of training samples while preserving previously learned class knowledge [25], [31], [32]. This technology has attracted widespread attention and spurred the development of various algorithms, which can be divided into two main types: one uses a "Feature Extractor + Softmax Classifier" approach, where the entire network is trainable throughout the incremental learning process. To counteract catastrophic forgetting, it often incorporates additional mechanisms to consolidate old knowledge while learning new information, such as data replay and knowledge distillation. Another employs a "Feature Embedding + Nearest Mean Classifier" strategy, focusing on training an embedding network that maps samples into a feature space to highlight semantic differences, followed by classification using the nearest mean classifier. This typically involves learning an effective backbone, averaging the feature embeddings of the training data to serve as prototypes for the respective classes, and using metrics like cosine similarity for predicting testing data. Several representative methods [33]–[40] have been developed for FSCIL. The CEC algorithm [34] employs a graph attention network to update relationships between base and new prototypes, aiding classifiers in finding more precise decision boundaries in complex data environments. The FACT framework [35] introduces the concept of forward compatibility to enhance the model's adaptability to future incremental classes. Although these technologies have begun to be applied in other medical diagnostic areas such as skin disease classification [41], there has yet to be research exploring the application of FSCIL in retinal disease diagnosis.

C. Foundation Models

Recently, the rise of foundation models has sparked extensive discussions, particularly with the emergence of vision-language models and vision-based models. These models have achieved significant success in traditional computer vision and natural language processing tasks and have excelled in cross-modal learning and multimodal tasks. Recently, a series of works [42]–[45] have applied CLIP [46] to general FSCIL problems, leveraging approaches such as CA-CLIP proposed by *Xu et al.* [42], PL-FSCIL proposed by *Tian et al.* [43], and CPE-CLIP proposed by *D'Alessandro et al.* [44]. However, despite CLIP's effectiveness in general FSCIL tasks, its lack of specialization for retinal image analysis poses potential challenges in the specific domain of retinal disease diagnosis. Recently, a retinal image model named RETFound [11] has gained widespread attention. This model utilizes large-scale unlabeled retinal images for self-supervised learning, thereby acquiring universal retinal representations. Subsequently, the model is fine-tuned on tasks with explicit labels, resulting in significant performance improvements. Therefore, considering its superior performance in specialized domains, we plan to introduce the RETFound model into the problem of FSCIRDR to enhance the model's generalization ability and adaptability.

III. Method

In this section, we first define FSCIL's problem setting. Then, we analyze FSCIRDR's challenges. We then outline the proposed framework and provide a comprehensive breakdown of its components.

A. Problem Setting

Consider $\{D^0_{train}, \cdots, D^n_{train}\}$ and $\{D^0_{test}, \cdots, D^n_{test}\}$ as the training and testing datasets for the FSCIL, respectively. Here, n represents the number of incremental sessions within the current FSCIL task. D^0_{train} is the training dataset for the base session, which includes a substantial amount of labeled training data. For any integer i from 1 to n, D^i_{train} adopts an $N-$way $K-$shot format, meaning that the training dataset for session i comprises N classes, each with K labeled samples. D^i_{test} represents the testing dataset for session i. For any integers i, j ranging from 0 to n where $i \neq j$, the corresponding label spaces of D^i_{train} and D^i_{test}, denoted as C^i, do not intersect, *i.e.*, $C^i \cap C^j = \emptyset$. When the learning

process progresses to session i, only D^i_{train} is accessible, while the complete training datasets from previous sessions are unavailable. The evaluation for session i includes the testing datasets from all previous and current sessions, denoted as $D^0_{test} \cup \cdots \cup D^i_{test}$.

B. Challenge Analysis

In this study, we examine the principal challenges encountered in FSCIRDR, which include the overfitting problem due to limited training data, the catastrophic forgetting during incremental learning, and the fine-grained challenge among retinal disease classes.

1) Overfitting: In FSCIRDR, limited data availability often directs the model training efforts toward minimizing prediction errors on the training dataset. This approach is particularly prone to significant discrepancies between empirical and expected risks when the dataset adopts an $N-$way $K-$shot format, leading to the overfitting problem where the model performs well on training data but fails on test data. Furthermore, as the new classes are incrementally added, continuous reliance on this unreliable empirical risk minimization strategy may hinder the model from achieving an ideal state, challenging the model's reliability and stability in current and subsequent sessions.

2) Catastrophic Forgetting: FSCIRDR demands a balance between maintaining stability for previously learned knowledge and exhibiting plasticity for new classes. Indiscriminate optimization of existing model parameters when introducing new classes could lead to decision boundaries biased towards new classes, resulting in catastrophic forgetting. Conversely, a moderate focus on the stability of old knowledge could allow the model's ability to learn new tasks.

3) Fine-grained Challenges: When dealing with various retinal diseases, the visual differences between the classes are often subtle, requiring precise discriminative abilities from the model. Introducing new classes similar to existing ones can make classification more difficult and confuse the model between old and new classes. Moreover, having limited data for new classes and their strong resemblance to old ones makes it challenging for the model to accurately differentiate and adapt to these new, fine-grained classes.

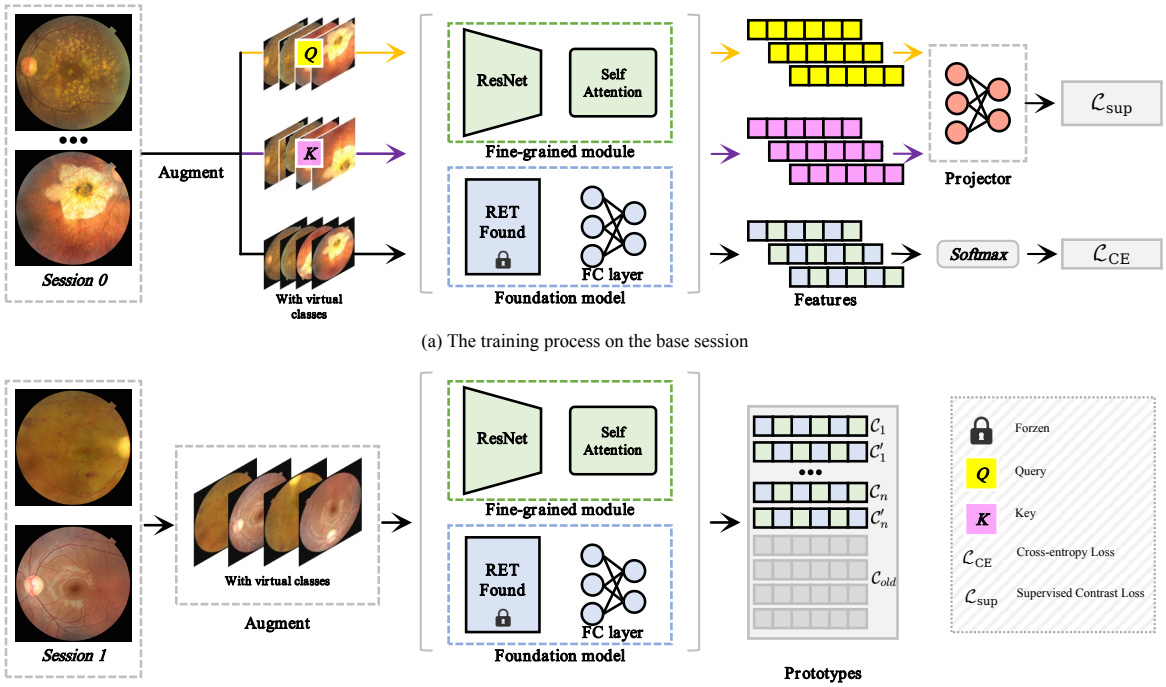
C. Our Framework

In response to the aforementioned challenges, we introduce a specialized FSCIL framework for retinal disease recognition, Re-FSCIL, as depicted in Fig. 2. This framework integrates the foundational RETFound model with a fine-grained module. Based on existing related methods [25], [34], [47], [48], which indicate that the "Feature Embedding + Nearest Mean Classifier" strategy is currently a cost-effective approach for FSCIL settings, our Re-FSCIL also adopts this strategy. During the base session, the model is initialized, and in subsequent incremental sessions, the parameters are frozen to extract class prototypes and classify testing samples using the nearest mean classifier. Since class prototypes are obtained by averaging the feature embeddings of each class, there is no need to retrain the model. This method helps avoid overfitting caused by retraining with limited samples and mitigates catastrophic forgetting to some extent. This strategy does not rely on data replay and knowledge distillation when handling incremental sessions, thereby avoiding memory requirements, privacy risks, and potential data imbalance issues. However, directly using this strategy may limit the model's ability to adapt to future new classes, especially when there is a high degree of fine-grained similarity between new and old classes.

To address these concerns, we employ a series of strategies. Specifically, the training process is divided into two main stages: base training and prototype generation. The base training comprises three pivotal components: first, forward compatibility training of the fine-grained module, emphasizing forward compatibility by simulating potential future classes to provide a foresight perspective; second, enhancement of the model's discriminative ability through supervised contrastive learning, focusing on improving the capability to differentiate subtle features; and third, integration of pre-trained knowledge from the foundational model to bolster feature representation. Detailed descriptions of these components are described in subsequent sections. During the prototype generation phase, the model parameters established in the base session are fixed, and features are extracted and averaged from each class's training samples to formulate class prototypes. These prototypes subsequently act as weights for a cosine similarity-based classifier, facilitating accurate classification.

1) Forward-compatible Fine-grained Module: As shown in Fig. 2, our forward-compatible fine-grained module comprises ResNet and self-attention. In conventional visual tasks, the high-level features learned by deep neural networks can effectively handle coarse-grained tasks. However, in the field of retinal disease classification, distinguishing between diseases is a significant challenge due to the subtle visual differences between disease types, which can be regarded as a fine-grained recognition problem. Additionally, since FSCIRDR involves multiple incremental stages, it falls under the category of long-sequence tasks. Each stage involves completely independent retinal disease classes, requiring the model to have good fine-grained discrimination ability and to adapt to new disease recognition tasks. Under these conditions, many attention mechanisms need to be adjusted for specific tasks, whereas the self-attention model can adapt well to long-sequence and complex dependency tasks. Besides, considering that the self-attention mechanism has been widely applied and has shown excellent results in fine-grained recognition and FSCIL tasks, we introduced the self-attention mechanism to enhance the model's ability to learn fine-grained feature representations.

Due to the adoption of the "Feature Embedding + Nearest Mean Classifier" strategy to handle FSCIRDR, our framework's ability to adapt to future changes is somewhat limited. To address this, we provide more forward-looking perspectives for our framework during the base session training, enabling the model to handle new classes in the future. Specifically, we employed a forward-compatible strategy by creating virtual classes to simulate potential future incremental classes. These virtual classes, generated using rotation augmentation, are combined with the real base classes to finish the training of our framework. The generation process can be expressed



(a) The training process on the base session

(b) Prototypes generation process after base session training

Fig. 2: Our proposed Re-FSCIL framework. (a) shows the training process in the base session. Initially, training samples are augmented into various forms required by different modules. Subsequently, augmented images with virtual classes are fed into the encoder to obtain initial features. Then, the features of ResNet processed by the self-attention module are fused with the features of RETFound after dimensionality reduction. The framework is then optimized using cross-entropy loss. Moreover, the augmented key and query images are utilized for supervised contrast loss to optimize the model further. (b) illustrates the process of generating class prototypes. After completing the base training, the model's parameters are fixed, and class prototypes are calculated by averaging the features of samples from each class, which are then used for subsequent classification.

as $\mathcal{F}(x, y) = \{(x_m, y_m)\}_{m=1}^{M}$, where (x_m, y_m) denotes the m-th transformation applied to (x, y), and $\mathcal{F}$ represents the transformation function. Note that (x_1, y_1) represents the original image-label pair and $M = 2$ in this paper. The virtual class generation can enrich the training samples, act as placeholders in the feature space for future updates, and help compress existing class distributions in the feature space to enhance discrimination. This approach ensures that the model can handle current tasks and is also well-equipped for upcoming classes, significantly improving its performance in scenarios requiring fine-grained FSCIL.

2) Fusion with Foundation Model Feature: Although the Re-FSCIL framework adopts the "Feature Embedding + Nearest Mean Classifier" strategy for classification, the fine-grained module, which completes parameter initialization based only on the training data from the base session, may not ensure that the model learns sufficiently discriminative and representative features for new classes. While the forward-compatible strategy has been implemented to improve adaptability to incremental classes to some extent, this is not entirely adequate.

Recent studies have shown that foundation models pre-trained on extensive datasets can exhibit excellent performance on certain classification tasks without any training samples. Thus, we consider incorporating knowledge from foundation models to enhance our framework's overall feature extraction capability. Since general foundation models may not be well-suited to specialized analytic tasks such as retinal disease recognition, we propose integrating the RETFound model, which has been pre-trained on a large-scale dataset of retinal images. RETFound employs the self-supervised learning approach, *Masked Autoencoder* (MAE), to pre-train the large vision Transformer on 1.6 million unlabeled retinal images. It has demonstrated good adaptability in many downstream tasks. The core idea of MAE is to mask a portion of the pixels in an image and then train the model to reconstruct the masked parts, significantly enhancing the model's understanding and representation of images.

In our framework, we use a fully connected layer to reduce the dimensionality of the features extracted by RETFound and fuse them with the features extracted by our fine-grained module. Initially, combined with the virtual classes, our training for our framework guided by a simple cross-entropy loss function can be expressed as:

$$\mathcal{L}_{total}\left(\phi; x_m, y_m\right) = \mathcal{L}_{ce}\left(\phi\left(x_m\right), y_m\right), \tag{1}$$

where $\mathcal{L}_{ce}\left(\cdot\right)$ denotes the CE loss, x represents the sample, y is the corresponding label, and the model $\phi\left(\cdot\right)$ consists of our fine-grained module and RETFound feature extractor, which can be expressed as:

$$\phi\left(\cdot\right) = W^T g\left(\cdot\right) = W^T\left(\left(1-\alpha\right) f_{fg}\left(\cdot\right) + \alpha f_{re}\left(\cdot\right)\right), \tag{2}$$

where W represents the classifier, $f_{fg}\left(\cdot\right)$ is the fine-grained

module, $f_{re}(\cdot)$ denotes the RETFound feature extractor, and α is the weight for fusing the features. It is important to note that the testing data used in our experiments are independent of the training data used by RETFound.

3) Supervised Contrast Learning: Although the introduction of the self-attention mechanism improves the model's ability to discriminate different retinal diseases to some extent, it is not sufficient to train the fine-grained module by conventional cross-entropy loss. Therefore, in order to further improve the model's ability to identify different retinal diseases, we introduce supervised contrast loss $\mathcal{L}_{sup}$, which is designed to minimize intra-class differences while maximizing inter-class separation, thereby enhancing the model's performance in classifying different retinal diseases. Specifically, given a batch of image-label pairs $\{(x_i, y_i)\}_{i=0}^{b}$, where each image undergoes random augmentations to generate a query view $x_q = Aug_q(x)$ and a key view $x_k = Aug_k(x)$, these views are then fed in $\psi(\cdot)$ to get L_2-normalized representations q and k, where $\psi = h \circ g$ is composed of the entire image encoder g and a projector h. The supervised contrastive loss $\mathcal{L}_{sup}$ is calculated using these representations to optimize the model's feature discrimination capabilities, formulated as:

$$\mathcal{L}_{sup}(\psi; x_i, y_i, \mathcal{T}) = -\frac{1}{|\mathbf{k}_+|} \sum_{k_+ \in \mathbf{k}_+} \log \frac{\exp(q_i^\mathsf{T} k_+/\mathcal{T})}{\sum_{k \in \mathbf{k}} \exp(q_i^\mathsf{T} k/\mathcal{T})}, \quad (3)$$

where $\mathbf{k}$ is the set of all the key representations, $\mathbf{k}_+$ denotes positive set, *i.e.*, those in $\mathbf{k}$ belonging to the same class with x_i, and the $\mathcal{T}$ is a temperature parameter. After the introduction of the supervised contrastive loss, the joint training loss can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \beta \mathcal{L}_{sup}, \quad (4)$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss function, $\mathcal{L}_{sup}$ represents the supervised contrast loss function, and β is a hyper-parameter that weights the importance of $\mathcal{L}_{sup}$.

4) Prototype Generation and Inference: After training in the base session, the model's parameters are fixed to generate the prototypes in the classifier. As shown in Fig. 2, for session I, each class is augmented to generate a corresponding virtual class. The fixed model is then used to calculate the prototypes by averaging the feature embeddings extracted from both the real and virtual classes. These prototypes are concatenated with the previous ones in the classifier. The process can be described as:

$$W = \bigcup_{i=0}^{I} \left\{ \boldsymbol{w}_{cm}^i \mid 1 \le c \le |\mathcal{C}^i|, 1 \le m \le M \right\}, \quad (5)$$

where $\boldsymbol{w}_{ij}^t$ represents the prototypes. During the inference, we use the same method to augment the test sample: $\mathcal{F}(x) = \{x_m\}_{m=1}^{M}$. Then, we aggregate the inference results based on the predictions for both the real and virtual samples:

$$c_x = \arg\max_{c,i} \sum_{m=1}^{M} \mathrm{sim}\left(g(x_m), \boldsymbol{w}_{cm}^i\right), \quad (6)$$

where $\mathrm{sim}(\mathbf{p}, \mathbf{q}) = \mathbf{p}^\mathsf{T}\mathbf{q}/(\|\mathbf{p}\|\|\mathbf{q}\|)$ is cosine similarity between two vectors.

IV. Experiments

A. Dataset and Metric

RFMiD38: The RFMiD dataset [1] is a publicly available dataset designed to detect multiple retinal diseases, containing 46 different disease conditions. Given the uneven sample distribution across classes in this dataset, we have selected 38 classes to construct a new benchmark for FSCIRDR. We use 20 classes as base classes and 18 as incremental classes. These 18 incremental classes are divided into six incremental sessions, each covering three classes with three training samples per class, thus establishing a $3-$way $3-$shot configuration. Examples of some samples are shown in Fig. 3.

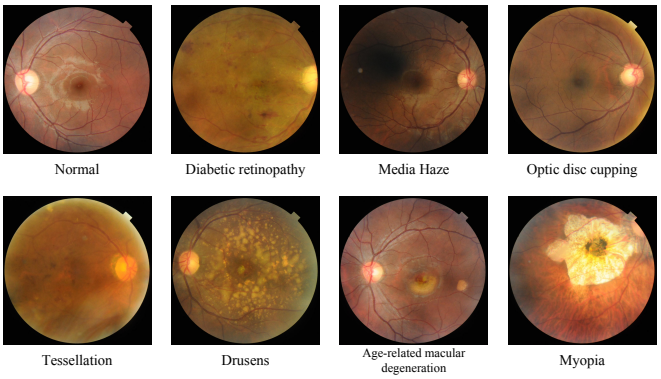


Fig. 3: Examples in the RFMiD38 dataset.

JSIEC39: The JSIEC dataset [49] is another publicly available dataset used for the detection of various retinal diseases. It contains 39 classes sourced from the Joint Shantou International Eye Centre, China. The number of samples varies across different classes. We use 21 classes as base classes and 18 as incremental classes. These 18 incremental classes are further divided into 6 incremental sessions, each covering 3 classes with 3 training samples per class, thereby establishing a $3-$way $3-$shot configuration. Examples of some samples are shown in Fig. 4.

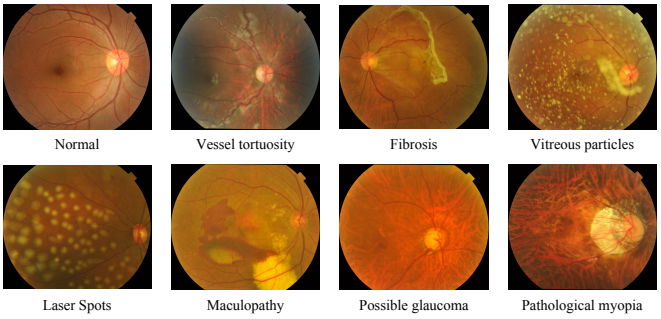


Fig. 4: Examples in the JSIEC39 dataset.

To comprehensively evaluate our framework, we utilized three metrics: 1) Accuracy values for each session; 2) The *Performance Drop* (PD) rate [34], which measures the absolute decline in accuracy from the initial session to the final session; 3) The *Average Accuracy* (AA) across all sessions. For accuracy and AA, higher values indicate better performance, while for PD, lower values are preferable.

B. Implementation Details

In our framework, we use the RETFound model, 'vit_large_patch16," which is a foundational model pre-trained on a large-scale retinal dataset. We keep the parameters of the RETFound model fixed throughout and use it solely for feature extraction. The fine-grained module adopts ResNet-18 [50], and we use the pre-trained weights provided by PyTorch for initialization. We use SGD with 0.9 momentum to optimize the model. The initial learning rate is 0.005 in the base session. The training epochs are 50. For augmentation operation Aug, only one rotation is used to generate virtual classes in this paper. Our framework is implemented in PyTorch 2.2 and Python 3.9 and trained on the Nvidia Tesla V100 GPU.

C. Comparison with State of The Arts

This study evaluates our method alongside several advanced methods on RFMiD38 and JSIEC39. Since we conducted a benchmark study, we provide detailed information about each compared method:

- **RETFound [11]:** It is a foundational model for retinal images that learns generalizable representations from 1.6 million unlabelled retinal images using self-supervised learning. It can be used for various downstream tasks in retinal disease diagnosis.
- **CEC [34]:** It tackles FSCIL by updating only the classifiers in each incremental session to avoid knowledge forgetting. It also introduces a continually evolved classifier that uses the graph attention network to propagate context information between classifiers for better adaptation.
- **FACT [35]:** It handles FSCIL by learning prospectively to prepare for future updates. It uses a forward-compatible training strategy to reserve embedding space for future new classes by assigning virtual prototypes, allowing the model to incorporate new classes efficiently while resisting forgetting old ones.
- **BiDist [51]:** It adapts KD for FSCIL using two teacher models: one trained on abundant base class data to reduce overfitting of novel classes, and the other from the last incremental session to alleviate forgetting. An adaptive strategy and a two-branch network with an attention-based aggregation module combine these guidances and preserve base knowledge.
- **SAVC [48]:** Similar to FACT, it adopts a forward-compatible strategy by introducing virtual classes to enhance supervised contrastive learning, facilitating the separation. These virtual classes act as placeholders for unseen classes in the representation space and provide diverse semantic information.
- **TEEN [36]:** It uses a training-free calibration strategy to enhance the discriminability of new classes by fusing the new prototypes with weighted base prototypes, thus improving the classification performance of new classes.

Tab. I and Tab. II provide detailed accuracy performance for each method in different sessions on the RFMiD38 and JSIEC39 datasets, also including AA and PD. Additionally, to demonstrate the performance variations of each method across different sessions, we have plotted the performance in Fig. 5. Notably, there is a significant performance disparity between these two datasets for all methods, indicating the substantial impact of dataset characteristics on performance.

TABLE I: Comparison with SOTA methods on RFMiD38. We implemented the results of the compared methods on the RFMiD38 dataset using the officially published code. (In %)

Method	Venue	Acc. in each session							AA↑	PD↓
		0	1	2	3	4	5	6		
RETFound	*Nature23*	52.09	36.41	29.38	33.22	30.10	28.09	25.28	33.51	26.81
CEC	*CVPR21*	28.12	27.33	22.78	22.56	22.01	22.08	18.56	23.35	9.56
FACT	*CVPR22*	21.42	16.27	12.60	15.80	13.39	11.97	9.17	14.37	12.25
BiDist	*CVPR23*	21.23	19.85	18.94	16.89	15.66	14.60	13.21	17.20	8.02
SAVC	*CVPR23*	58.11	56.27	46.75	46.93	45.83	44.92	37.89	48.10	20.22
TEEN	*NIPS24*	18.05	11.77	9.40	10.07	9.07	8.88	8.11	10.76	9.94
Re-FSCIL	*Ours*	**64.83**	**62.32**	**53.51**	**53.86**	**51.69**	**49.41**	**42.72**	**54.05**	22.11

TABLE II: Comparison with SOTA methods on JSIEC39. We implemented the results of the compared methods on the JSIEC39 dataset using the officially published code. (In %)

Method	Venue	Acc. in each session							AA↑	PD↓
		0	1	2	3	4	5	6		
RETFound	*Nature 23*	73.53	66.25	62.39	57.95	58.74	55.20	51.67	60.82	21.87
CEC	*CVPR21*	47.56	46.82	43.34	39.10	37.39	28.80	32.20	39.32	15.36
FACT	*CVPR22*	24.57	21.02	19.05	17.31	18.66	13.40	15.40	18.49	9.17
BiDist	*CVPR23*	31.50	28.33	27.33	26.65	24.87	22.64	20.47	25.97	11.03
SAVC	*CVPR23*	85.14	81.52	75.91	74.30	73.35	67.4	65.93	74.79	19.21
TEEN	*NIPS24*	26.71	24.27	20.66	19.55	20.47	15.00	16.33	20.43	10.38
Re-FSCIL	*Ours*	**90.63**	**85.93**	**76.39**	**75.78**	**77.54**	**71.60**	**71.40**	**78.47**	19.23

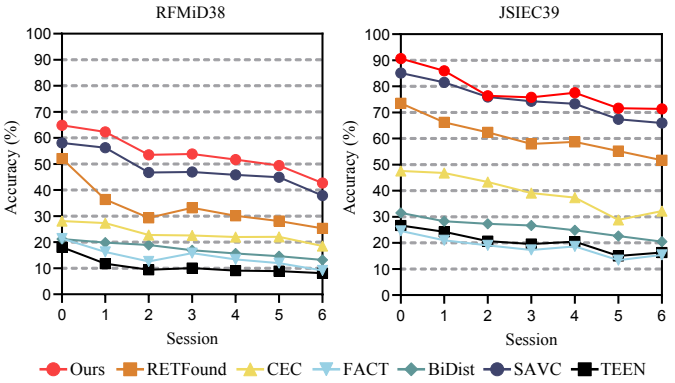


Fig. 5: Comparison with SOTA methods on RFMiD38 and JSIEC39 datasets. Our method significantly surpasses existing advanced methods.

Our Re-FSCIL method demonstrates superior performance on both the RFMiD38 and JSIEC39 datasets. Specifically, on the RFMiD38 dataset, Re-FSCIL achieves an AA of 54.05% across all sessions, significantly outperforming other methods such as RETFound (33.51%), CEC (23.35%), FACT (14.37%), BiDist (17.20%), SAVC (48.10%), and TEEN (10.76%). In terms of PD, our method obtains a PD value of 22.11% on the RFMiD38 dataset. In each session, Re-FSCIL's performance is

outstanding. For instance, in session 0, Re-FSCIL's accuracy is 64.83%, significantly higher than the best-performing alternative method SAVC (58.11%). In session 1, Re-FSCIL achieves an accuracy of 62.32%, also notably higher than the second-best method, SAVC (56.27%). Re-FSCIL maintains its leading accuracy on subsequent sessions, such as achieving 53.86% in session 3, compared to SAVC's 46.93%.

Similarly, on the JSIEC39 dataset, the Re-FSCIL method performs excellently. Our method achieves an AA of 78.47%, far surpassing other methods such as RETFound (60.82%), CEC (39.32%), FACT (18.49%), BiDist (25.97%), SAVC (74.79%), and TEEN (20.43%). In each session on the JSIEC39 dataset, Re-FSCIL continues to demonstrate outstanding performance. For example, in session 0, Re-FSCIL achieves an accuracy of 90.63%, far exceeding the best alternative method SAVC (85.14%). In session 1, Re-FSCIL's accuracy is 85.93%, significantly higher than the second-best method SAVC (81.52%). In subsequent sessions, Re-FSCIL maintains its superior accuracy, such as 75.78% in session 3, compared to SAVC's 74.30%.

To illustrate the specific classification performance of our proposed method on each base and incremental class, refer to the confusion matrices provided in Fig. 6(d) and Fig. 6(h), which show the model's performance on RFMiD38 and JSIEC39 datasets in the last incremental session. Our method effectively classifies both base and incremental classes, demonstrating the ability to address potential overfitting issues, catastrophic forgetting, and fine-grained challenges.

In summary, our Re-FSCIL method surpasses existing methods on both the RFMiD38 and JSIEC39 datasets in terms of accuracy and stability. These outstanding results indicate that our method has significant advantages and reliability in handling complex situations and adapting to various scenarios. However, the performance differences between datasets highlight the crucial impact of dataset characteristics on algorithm performance. Specifically, factors such as dataset size and quality, annotation accuracy and consistency, class distribution and imbalance, disease characteristics and complexity, and image acquisition methods and conditions can all affect model performance. Therefore, when evaluating algorithms, it is essential to consider the diversity of dataset characteristics to fully reflect the algorithm's applicability and stability. This reminds us that in practical applications, algorithm selection and adjustments should be made based on the specific characteristics of the dataset to achieve optimal performance.

D. Ablation Study

To substantiate the significance of our proposed components, we conducted ablation studies focusing on the key aspects of our method. These components include the foundation model, the fine-grained module, supervised contrastive learning, and a forward-compatible strategy. We report the results for RFMiD38 in Tab. III and JSIEC39 in Tab. IV, demonstrating the impact of each component on the overall performance.

Compared to the initial baseline model RETFound, the integration of a fine-grained module has resulted in significant AA improvements on the RFMiD38 and JSIEC39 datasets, achieving increases of 9.75% and 6.93%, respectively. This enhancement enables our framework to distinguish between fine-grained retinal disease classes effectively. Supervised contrastive learning further improves the model's classification capability for fine-grained classes, enhancing its understanding of inter-class differences. This approach brings additional AA gains of 2.73% and 6.97% on the RFMiD38 and JSIEC39 datasets, respectively. However, we observed that not all incremental sessions showed performance improvements. This is because supervised contrastive learning focuses on enhancing the model's ability to discriminate between classes in the base sessions. This phenomenon can also be seen in the confusion matrices in Fig. 6, where the accuracy for base session classes increases. However, the performance in incremental sessions remains mediocre. This does not indicate instability in the model's performance but rather reflects the different emphases of each module. To further enhance generalization capabilities, we introduced a forward-compatible strategy after supervised contrastive learning. This strategy generates virtual classes to simulate potential future classes, thereby enhancing the model's generalization capability. As a result, this approach leads to further AA improvements of 8.06% and 3.75% on the RFMiD38 and JSIEC39 datasets, respectively.

TABLE III: Ablation study on RFMiD38. For short, **FM** stands for the foundation model; **FG** stands for the integration of fine-grained module; **SC** represents the supervised contrastive learning; **FC** indicates the forward-compatible strategy. (In %)

FM	FG	SC	FC	Acc. in each session							AA↑	PD↓
				0	1	2	3	4	5	6		
✓				52.09	36.41	29.38	33.22	30.10	28.09	25.28	33.51	26.81
✓	✓			54.31	52.00	43.53	41.67	40.27	38.48	32.56	43.26	21.76
✓	✓	✓		57.89	51.72	43.35	45.40	44.25	42.46	36.89	45.99	21.00
✓	✓	✓	✓	64.83	62.32	53.51	53.86	51.69	49.41	42.72	54.05	22.11

TABLE IV: Ablation study on JSIEC39. For short, **FM** stands for the foundation model; **FG** stands for the integration of fine-grained module; **SC** represents the supervised contrastive learning; **FC** indicates the forward-compatible strategy. (In %)

FM	FG	SC	F	Acc. in each session							AA↑	PD↓
				0	1	2	3	4	5	6		
✓				73.53	66.25	62.39	57.95	58.74	55.20	51.67	60.82	21.87
✓	✓			77.11	73.61	66.25	63.44	63.97	70.60	59.27	67.75	17.85
✓	✓	✓		85.77	80.07	75.16	72.28	71.76	75.40	62.60	74.72	23.17
✓	✓	✓	✓	90.63	85.93	76.39	75.78	77.54	71.60	71.40	78.47	19.23

E. Visualization

To further investigate the contributions of each component of our framework, we display the confusion matrices generated by models in the last incremental session of our ablation studies on the RFMiD38 and JSIEC39 datasets in Fig. 6. Redder diagonals indicate higher classification accuracy against a dim background. Our observations reveal that the baseline model RETFound performs poorly on these datasets. However, with

the integration of different components, the model exhibits a noticeable improvement in performance for both new and old classes. This demonstrates that our method effectively adapts to new classes and accurately recognizes old classes, avoiding confusion in established decision boundaries.

To validate the effectiveness of our method in classifying classes in FSCIRDR, we visualized the feature space of the JSIEC39 dataset using t-SNE in the final session and roughly drew decision boundaries using the *Support Vector Machine* (SVM) (please note that these are not the actual decision boundaries, but are shown for better illustration), as shown in Fig. 7. To investigate each component of our framework's contribution further, we randomly selected 4 base classes and 2 incremental classes and evaluated the separation degree in the feature space of the original RETFound model and the model with our components incrementally added. The red circles in the figure indicate poorly separated classes. It can be observed that our complete framework significantly enhances the separation of previously poorly separated classes. These findings underscore the superior discriminative capability of our framework in addressing the challenges of FSCIRDR.

F. Impact of Hyper-parameter

In our framework, key hyperparameters include the balance between our fine-grained module and RETFound features, denoted by α in Eq. 2, and the weight of the supervised contrastive loss, β in Eq. 4. To thoroughly assess the impact of these hyperparameters on model performance, we show the performance achieved in each session after the final session learning with different parameter settings on the RFMiD38 and JSIEC39 datasets in Fig. 8. Observations indicate that the optimal hyperparameter settings for achieving the best performance on each dataset are: for RFMiD38, $\{\alpha, \beta\} = \{0.1, 0.5\}$; for JSIEC39, $\{\alpha, \beta\} = \{0.2, 0.5\}$.

V. Conclusion

This paper introduced the Re-FSCIL framework for FSCIRDR, addressing several challenges in retinal disease diagnosis. Our framework integrates the RETFound model with a fine-grained module, incorporating forward-compatible training, supervised contrastive learning, and feature fusion to improve model adaptability, feature discrimination, and representation quality. We converted existing datasets into the FSCIL format and reproduced numerous representative FSCIL methods, establishing two new benchmarks (RFMiD38 and JSIEC39) for FSCIRDR. We conducted comprehensive experiments, including comparisons with advanced methods, ablation studies, and hyperparameter analysis. Our experimental results indicate that Re-FSCIL outperforms existing methods on these benchmarks, representing a promising approach for few-shot continuous learning in retinal disease classification. However, there remains a significant performance disparity between different datasets. Although our method performs well on both datasets, it shows notably better performance on JSIEC39 compared to RFMiD38. This discrepancy may stem from differences in dataset characteristics, such as sample distribution and data complexity. This indicates that even the same method can exhibit significant performance variations across different datasets, highlighting the need to consider the diversity of dataset characteristics when designing and evaluating algorithms.

Despite the excellent performance of our method in experiments, this study has some limitations that need further discussion. The "Feature Embedding + Nearest Mean Classifier" strategy works well for short-term incremental tasks but has limitations for long-term incremental learning tasks. The increase in the number of classes and the complexity of the feature space may lead to performance degradation, suggesting that the base model should be retrained after accumulating a certain number of classes. Additionally, our method faces challenges in practical applications, especially due to the inconsistency of imaging devices and data quality. The high cost of acquiring and annotating retinal disease samples, along with the variations in imaging devices and data quality, can increase the complexity of practical applications. Therefore, appropriate adjustments to the model are necessary based on specific application conditions to improve its robustness and reliability. This highlights the need for further research to address performance differences across datasets and to optimize the algorithm's generalization capability, ensuring stable performance across various datasets.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 62306323 and the China Scholarship Council under Grant 202206110005. Jinghua Zhang and Peng Zhao contribute equally. Dewen Hu and Chen Li are the corresponding authors.

[1] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, L. Giancardo, G. Quellec, and F. Mériaudeau, "Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, 2021.

[2] M. H. Sarhan, M. A. Nasseri, D. Zapp, M. Maier, C. P. Lohmann, N. Navab, and A. Eslami, "Machine learning techniques for ophthalmic data processing: a review," *IEEE JBHI*, vol. 24, no. 12, pp. 3338–3350, 2020.

[3] M. A. Rodríguez, H. AlMarzouqi, and P. Liatsis, "Multi-label retinal disease classification using transformers," *IEEE JBHI*, 2022.

[4] K. Mittal and V. M. A. Rajam, "Computerized retinal image analysis-a survey," *Multimedia tools and Applications*, vol. 79, no. 31, pp. 22 389–22 421, 2020.

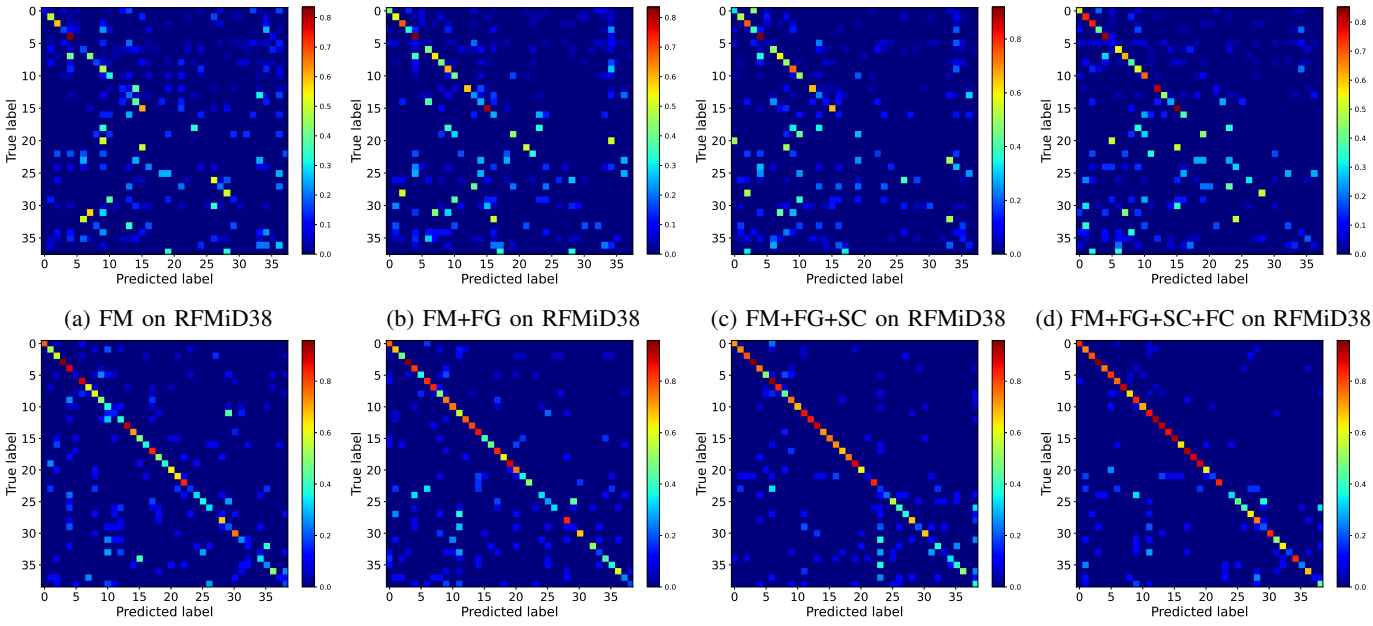
[5] J. Wen, D. Liu, Q. Wu, L. Zhao, W. C. Iao, and H. Lin, "Retinal image-based artificial intelligence in detecting and predicting kidney diseases: Current advances and future perspectives," *View*, vol. 4, no. 3, p. 20220070, 2023.

[6] Y. Y. Tan, H. G. Kang, C. J. Lee, S. S. Kim, S. Park, S. Thakur, Z. Da Soh, Y. Cho, Q. Peng, Y.-C. Tham *et al.*, "Prognostic potentials of ai in ophthalmology: systemic disease forecasting via retinal imaging," *Eye and Vision*, vol. 11, no. 1, p. 17, 2024.

[7] W.-H. Yang, B. Zheng, M.-N. Wu, S.-J. Zhu, F.-Q. Fei, M. Weng, X. Zhang, and P.-R. Lu, "An evaluation system of fundus photograph-based intelligent diagnostic technology for diabetic retinopathy and applicability for research," *Diabetes Therapy*, vol. 10, pp. 1811–1822, 2019.

[8] X. Liu and W. Chi, "A cross-lesion attention network for accurate diabetic retinopathy grading with fundus images," *IEEE TIM*, 2023.

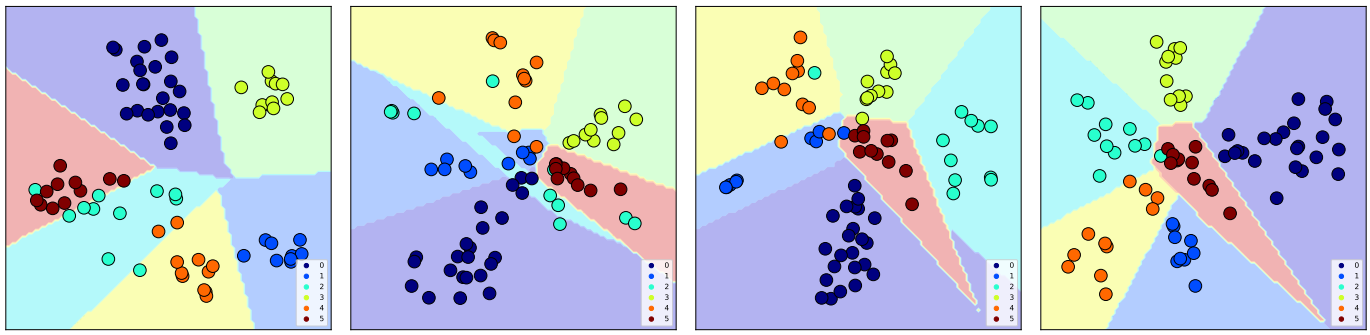
[9] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D. I. Fotiadis, and K. Marias, "Deep learning for diabetic retinopathy detection and classification based on fundus images: A review," *Computers in biology and medicine*, vol. 135, p. 104599, 2021.



(a) FM on RFMiD38 (b) FM+FG on RFMiD38 (c) FM+FG+SC on RFMiD38 (d) FM+FG+SC+FC on RFMiD38

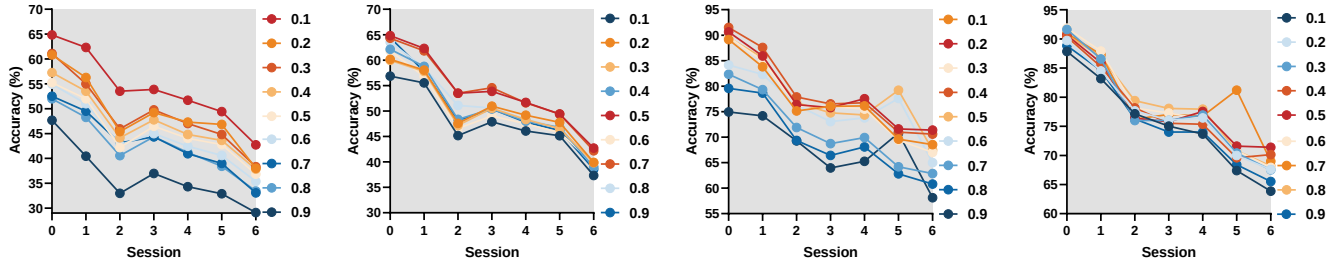
(e) FM on JSIEC39 (f) FM+FG on JSIEC39 (g) FM+FG+SC on JSIEC39 (h) FM+FG+SC+FC on JSIEC39

Fig. 6: Comparison of the confusion matrices of different ablation methods on RFMiD38 and JSIEC39 datasets. Redder diagonals indicate higher classification accuracy against a dim background. **FM** stands for the foundation model; **FG** stands for the integration of fine-grained module; **SC** represents the supervised contrastive learning; **FC** indicates the forward-compatible strategy.



(a) FM (b) FM+FG (c) FM+FG+SC (d) FM+FG+SC+FC

Fig. 7: The t-SNE visualization of the features learned by different ablation methods on the JSIEC39 dataset. Classes 0-3 represent the base classes, while classes 4-5 represent the incremental classes. Different background colors show different decision binaries. It can be found that our complete method gets the best class separation degree.



(a) The α influence on RFMiD38 (b) The β influence on RFMiD38 (c) The α influence on JSIEC39 (d) The β influence on JSIEC39

Fig. 8: Hyper-parameter influence on RFMiD38 and JSIEC39 datasets.

[10] Y. Xie, Q. Wan, H. Xie, Y. Xu, T. Wang, S. Wang, and B. Lei, "Fundus image-label pairs synthesis and retinopathy screening via gans with class-imbalanced semi-supervised learning," *IEEE TMI*, 2023.

[11] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, 2023.

[12] M. Badar, M. Haris, and A. Fatima, "Application of deep learning for retinal image analysis: A review," *Computer Science Review*, vol. 35, p. 100203, 2020.

[13] A. Bourouis, M. Feham, M. A. Hossain, and L. Zhang, "An intelligent mobile based decision support system for retinal disease diagnosis," *Decision Support Systems*, vol. 59, pp. 341–350, 2014.

[14] M. S. Haleem, L. Han, J. van Hemert, B. Li, and A. Fleming, "Retinal area detector from scanning laser ophthalmoscope (slo) images for diagnosing retinal diseases," *IEEE JBHI*, vol. 19, no. 4, pp. 1472–1482, 2014.

[15] C. Muramatsu, T. Nakagawa, A. Sawada, Y. Hatanaka, T. Hara, T. Yamamoto, and H. Fujita, "Automated segmentation of optic disc region on retinal fundus photographs: Comparison of contour modeling and pixel classification methods," *Computer methods and programs in biomedicine*, vol. 101, no. 1, pp. 23–32, 2011.

[16] Q. Meng, L. Liao, and S. Satoh, "Weakly-supervised learning with complementary heatmap for retinal disease detection," *IEEE TMI*, vol. 41, no. 8, pp. 2067–2078, 2022.

[17] J. Hu, H. Wang, G. Wu, Z. Cao, L. Mou, Y. Zhao, and J. Zhang, "Multi-scale interactive network with artery/vein discriminator for retinal vessel classification," *IEEE JBHI*, vol. 26, no. 8, pp. 3896–3905, 2022.

[18] H. Raja, T. Hassan, M. U. Akram, and N. Werghi, "Clinically verified hybrid deep learning system for retinal ganglion cells aware grading of glaucomatous progression," *IEEE TBME*, vol. 68, no. 7, pp. 2140–2151, 2020.

[19] Q. Liu, H. Liu, Y. Zhao, and Y. Liang, "Dual-branch network with dual-sampling modulated dice loss for hard exudate segmentation in color fundus images," *IEEE JBHI*, vol. 26, no. 3, pp. 1091–1102, 2021.

[20] Z. Li, J. Zhang, T. Tan, X. Teng, X. Sun, H. Zhao, L. Liu, Y. Xiao, B. Lee, Y. Li *et al.*, "Deep learning methods for lung cancer segmentation in whole-slide histopathology images—the acdc@ lunghp challenge 2019," *IEEE JBHI*, vol. 25, no. 2, pp. 429–440, 2020.

[21] J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, and H. Li, "Lcu-net: A novel low-cost u-net for environmental microorganism image segmentation," *Pattern Recognition*, vol. 115, p. 107885, 2021.

[22] H. Jiang, Y. Yin, J. Zhang, W. Deng, and C. Li, "Deep learning for liver cancer histopathology image analysis: A comprehensive survey," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108436, 2024.

[23] M. M. Rahaman, E. K. Millar, and E. Meijering, "Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning," *Scientific Reports*, vol. 13, no. 1, p. 13604, 2023.

[24] H. Chang, B. Liu, Y. Zong, C. Lu, and X. Wang, "Eeg-based parkinson's disease recognition via attention-based sparse graph convolutional neural network," *IEEE JBHI*, 2023.

[25] J. Zhang, L. Liu, O. Silven, M. Pietikäinen, and D. Hu, "Few-shot class-incremental learning: A survey," *arXiv preprint arXiv:2308.06764*, 2023.

[26] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[27] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, "Automated identification and grading system of diabetic retinopathy using deep neural networks," *Knowledge-Based Systems*, vol. 175, pp. 12–25, 2019.

[28] L. Ju, Z. Yu, L. Wang, X. Zhao, X. Wang, P. Bonnington, and Z. Ge, "Hierarchical knowledge guided learning for real-world retinal disease recognition," *IEEE TMI*, 2023.

[29] Q. Zhou, H. Zou, and Z. Wang, "Long-tailed multi-label retinal diseases recognition via relational learning and knowledge distillation," in *MICCAI*. Springer, 2022, pp. 709–718.

[30] Q. Meng and S. Shin'ichi, "Adinet: Attribute driven incremental network for retinal image classification," in *CVPR*, 2020, pp. 4033–4042.

[31] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *CVPR*, 2020, pp. 12 183–12 192.

[32] S. Tian, L. Li, W. Li, H. Ran, X. Ning, and P. Tiwari, "A survey on few-shot class-incremental learning," *Neural Networks*, vol. 169, pp. 307–324, 2024.

[33] Z. Chi, L. Gu, H. Liu, Y. Wang, Y. Yu, and J. Tang, "Metafscil: A meta-learning approach for few-shot class incremental learning," in *CVPR*, 2022, pp. 14 166–14 175.

[34] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, "Few-shot incremental learning with continually evolved classifiers," in *CVPR*, 2021, pp. 12 455–12 464.

[35] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, L. Ma, S. Pu, and D.-C. Zhan, "Forward compatible few-shot class-incremental learning," in *CVPR*, 2022, pp. 9046–9056.

[36] Q.-W. Wang, D.-W. Zhou, Y.-K. Zhang, D.-C. Zhan, and H.-J. Ye, "Few-shot class-incremental learning via training-free prototype calibration," *NeurIPS*, vol. 36, 2024.

[37] L. Sun, M. Zhang, B. Wang, and P. Tiwari, "Few-shot class-incremental learning for medical time series classification," *IEEE JBHI*, 2023.

[38] Z. Ji, Z. Hou, X. Liu, Y. Pang, and X. Li, "Memorizing complementation network for few-shot class-incremental learning," *IEEE TIP*, vol. 32, pp. 937–948, 2023.

[39] Y. Tai, Y. Tan, S. Xiong, and J. Tian, "Mine-distill-prototypes for complete few-shot class-incremental learning in image classification," *IEEE TGRS*, vol. 61, pp. 1–13, 2023.

[40] B. Liu, B. Yang, L. Xie, R. Wang, Q. Tian, and Q. Ye, "Learnable distribution calibration for few-shot class-incremental learning," *IEEE TPAMI*, 2023.

[41] J. Xiao, J. Li, and H. Gao, "Fs3dciot: A few-shot incremental learning network for skin disease differential diagnosis in the consumer iot," *IEEE TCE*, 2023.

[42] Y. Xu, S. Huang, and H. Zhou, "Ca-clip: category-aware adaptation of clip model for few-shot class-incremental learning," *Multimedia Systems*, vol. 30, no. 3, pp. 1–14, 2024.

[43] S. Tian, L. Li, W. Li, H. Ran, L. Li, and X. Ning, "Pl-fscil: Harnessing the power of prompts for few-shot class-incremental learning," *arXiv preprint arXiv:2401.14807*, 2024.

[44] M. D'Alessandro, A. Alonso, E. Calabrés, and M. Galar, "Multimodal parameter-efficient few-shot class incremental learning," in *ICCV*, 2023, pp. 3393–3403.

[45] A. Kumar, C. Bharti, S. Dutta, S. Karanam, and B. Banerjee, "Few shot class incremental learning using vision-language models," *arXiv preprint arXiv:2405.01040*, 2024.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[47] H. Zhao, Y. Fu, M. Kang, Q. Tian, F. Wu, and X. Li, "Mgsvf: Multi-grained slow versus fast framework for few-shot class-incremental learning," *IEEE TPAMI*, vol. 46, no. 3, pp. 1576–1588, 2021.

[48] Z. Song, Y. Zhao, Y. Shi, P. Peng, L. Yuan, and Y. Tian, "Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning," in *CVPR*, 2023, pp. 24 183–24 192.

[49] L.-P. Cen, J. Ji, J.-W. Lin, S.-T. Ju, H.-J. Lin, T.-P. Li, Y. Wang, J.-F. Yang, Y.-F. Liu, S. Tan *et al.*, "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," *Nature communications*, vol. 12, no. 1, p. 4828, 2021.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[51] L. Zhao, J. Lu, Y. Xu, Z. Cheng, D. Guo, Y. Niu, and X. Fang, "Few-shot class-incremental learning via class-aware bilateral distillation," in *CVPR*, 2023, pp. 11 838–11 847.