Chinese Journal of Electronics
vol. 33, no. 1, pp. 139–152, January 2024
https://doi.org/10.23919/cje.2022.00.396

RESEARCH ARTICLE

FSCIL-EACA: Few-Shot Class-Incremental Learning Network Based on Embedding Augmentation and Classifier Adaptation for Image Classification

Ruru ZHANG[1,2], Haihong E[1,2], and Meina SONG[1,2]

1. *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China*
2. *Education Department Information Network Engineering Research Center, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Corresponding author: Meina SONG, Email: mnsong@bupt.edu.cn

Manuscript Received November 19, 2022; Accepted June 16, 2023

Abstract — The ability to learn incrementally is critical to the long-term operation of AI systems. Benefiting from the power of few-shot class-incremental learning (FSCIL), deep learning models can continuously recognize new classes with only a few samples. The difficulty is that limited instances of new classes will lead to overfitting and exacerbate the catastrophic forgetting of the old classes. Most previous works alleviate the above problems by imposing strong constraints on the model structure or parameters, but ignoring embedding network transferability and classifier adaptation (CA), failing to guarantee the efficient utilization of visual features and establishing relationships between old and new classes. In this paper, we propose a simple and novel approach from two perspectives: embedding bias and classifier bias. The method learns an embedding augmented (EA) network with cross-class transfer and class-specific discriminative abilities based on self-supervised learning and modulated attention to alleviate embedding bias. Based on the adaptive incremental classifier learning scheme to realize incremental learning capability, guiding the adaptive update of prototypes and feature embeddings to alleviate classifier bias. We conduct extensive experiments on two popular natural image datasets and two medical datasets. The experiments show that our method is significantly better than the baseline and achieves state-of-the-art results.

Keywords — Few-shot class-incremental learning, Embedding augmentation, Classifier adaptation, Image classification.

Citation — Ruru ZHANG, Haihong E, Meina SONG, "FSCIL-EACA: Few-Shot Class-Incremental Learning Network Based on Embedding Augmentation and Classifier Adaptation for Image Classification," *Chinese Journal of Electronics*, vol. 33, no. 1, pp. 139–152, 2024. doi: 10.23919/cje.2022.00.396.

I. Introduction

Deep learning techniques play an important role in image analysis tasks [1]–[3]. Most deep learning models are developed in closed scenarios with large-scale, high-quality datasets. However, actual task scenarios are usually dynamic and open, requiring a model to incrementally integrate new class knowledge without forgetting old classes [4], [5], which poses severe challenges to deep learning systems. Class incremental learning (CIL) [6]–[14] addresses this challenge to a certain extent when there are enough new class instances in the incremental task. However, many scenarios, such as the medical domain, suffer from data privacy and sparsity issues. First of all, due to the scarcity of medical data, the data for new diseases is often relatively small, which will cause severe overfitting problems and exacerbate the forgetting of old diseases. Secondly, due to privacy issues, when learning new disease classes, the old class data may no longer have the right to be used, so the overlap or confusion between the new and old class representations in the embedding space leads to catastrophic forgetting of the old diseases. These problems motivate research of few-shot class-incremental learning (FSCIL) without saving old

Associate Editor: Prof. Xun CAO, Nanjing University.

data, i.e., only access data of few-shot new classes in incremental tasks and learn a unified classifier that can recognize all visible classes.

The FSCIL research has only just begun. Current works mainly penalize parameters change by enforcing strong constraints on model structure [15] or model parameters [16]–[18] to mitigate catastrophic forgetting and overfitting. However, in [19], [20], a simple and effective Baseline model was found, which is trained only on the base class, and directly classifies old and new classes using the nearest class mean (NCM) classifier. It suffers from two major limitations: 1) Fixed feature extractors can retain the representation of learning, but ignore the representation of new classes, resulting in embedding bias. 2) When we directly add additional weight vectors of new classes in the incremental tasks, the discriminative decision boundary of the unified classifier may be severely biased. To alleviate the first problem, References [12], [13] and [17] introduced self-supervised learning (SSL) methods to mine more information to improve the quality of feature representations; Reference [14] learns transferable and diverse representations by letting the model see more classes during training through Mixup data augmentation. To alleviate the second problem, Reference [19] proposed a continually evolved classifier, which employs a graph model to propagate contextual information to update the classifier weights learned on each task.

Despite this, the above method ignores several important issues. 1) The number of new class samples is extremely limited, and it is difficult for traditional feature extraction networks to extract discriminative details, especially for fine-grained images, such as medical images. Therefore, how to make full use of image context information to improve the effective utilization of discriminative features is very important. 2) Gradually learning and adjusting old and new class prototypes also leads to a mismatch between fixed feature representations and classifier prototypes. Therefore, how to adaptively adjust the unified classifier and feature representation of old and new tasks is crucial for subsequent tasks.

In order to solve the above problems, we propose a new few-shot class-incremental learning based on embedding augmentation and classifier adaptation (FSCIL-EACA) model from the perspective of embedding learning and classifier learning. Specifically, in the embedded learning stage, we propose an embedding augmentation network (EAN). To learn the embedded features that can be migrated to new tasks, we introduced SSL into the network to improve the generalization of the model. To make better use of the unique distinguishing features of sparse samples, we added a modulated attention (MA) mechanism to obtain the weighted context information of each category, which is more conducive to extracting more representative features based on the global information of the image. To adaptively adjust the classifier weight and embedded features, we propose an adaptive incremental classifier (AIC) in the classifier learning stage to alleviate the classification bias of the FSCIL model. The process includes a hybrid relational projection (HRP) module and a pseudo-incremental episode selection (PES) module. HRP uses prototype self-projection (PSP) to establish a global context correlation between previous and current tasks to calibrate the weight of the unified classifier. Adaptively adjust the embeddings of the query set to adapt to the global classification task through query set cross-projection (QCP). At the same time, we hope that HRP has the learning ability of FSCIL incremental tasks, hence we propose a pseudo-incremental learning method based on meta-learning to conduct multi-stage training for HRP to quickly adapt to new tasks. Specifically, we build pseudo incremental tasks, namely meta tasks, through the PES module, and learn the generalization ability of the model between different meta tasks through the meta-learning mechanism. If the model can handle different types of simulated pseudo incremental tasks, it will easily handle the incoming “real” tasks with generality. Thus, when faced with a new and unprecedented task, it can also be better classified.

The main contributions of this paper are summarized as follows:

1) An embedding augmentation network is proposed to improve the generalization ability of the feature extraction network by introducing self-supervised learning, and a modulated attention mechanism is proposed to make full use of image context information to extract discriminative features of fine-grained sparse samples.

2) A novel and effective hybrid relational projection module is designed, which adaptively adjusts new and old class prototypes through prototype self-projection, and adjusts feature representations to match corresponding prototypes through query set cross-projection.

3) A pseudo-incremental episode selection module is proposed to obtain a rich meta-training set to simulate incremental learning tasks. Based on meta-learning training techniques, the hybrid relational projection module is endowed with the ability of continual learning.

4) Experiments on the CUB-200, MiniImagenet, HyperKvasir, and SKIN-7 datasets show that our method is significantly better than the baseline and sets a new state-of-the-art performance with a significant advantage.

II. Related Work

1. Class-incremental learning

Incremental learning aims to develop an artificial intelligence system that can process novel data continuously appearing in the real world while learning new knowledge, retaining or even integrating and optimizing old knowledge not to be forgotten. The current incremental learning methods mainly include:

1) Solutions based on rehearsal strategies. Rehearsal strategies [6]–[9], aim to store a limited sample of old classes to prevent forgetting previous tasks. Methods

such as iCaRL [6] and EEIL [7] learn to preserve the knowledge gained from the old class based on distillation losses. At present, there are many methods [10], [21], [22] to generate old data by training the GAN network, thereby avoiding the potential data privacy problems of rehearsal strategies. However, the generative model itself has not yet reached high accuracy, and the effect of this type of method is not satisfactory.

2) Solutions based on regularization. Regularization strategies such as elastic weight consolidation (EWC) [23], memory aware synapses (MAS) [24], and PathInt [25], aim to minimize the impact of essential weights on previous tasks when learning novel tasks. For example, EWC uses the Fisher information matrix to calculate the importance of network weights offline and slows down the learning of network weights that are highly relevant to previous tasks. PathInt calculates the integration strength of synapses online and expands it based on memory to accumulate information related to the task. However, as some works have noticed, these methods perform poorly in FSCIL scenarios.

3) Other solutions include NCM [11], which combines cosine normalization, forgetting constraints, and inter-class separation strategies to reduce the adverse effects of the imbalance between the previous data and the new data. This paper aims to reduce catastrophic forgetting in FSCIL without storing old data or using complex generative models. Reference [14] adopted explicit class augmentation and implicit semantic augmentation to address representation bias and classifier bias in class incremental learning, achieving state-of-the-art performance.

2. Few-shot learning

Few-shot learning (FSL) aims to adapt the model to recognize novel classes with very few samples, regardless of the model's performance in recognizing the base classes. Typical FSL algorithms need to extract task episodes from the overall data distribution for training. The data of each class in the episode is divided into a small support set and a more extensive query set. The number of classes in each episode is called "way", and the number of support images in each class is called "shot", so a group of five classes and one labeled image to form a "5way 1-shot" classification problem. Our work is more related to the FSL method based on meta-learning, which includes the metric-based learning method, optimization-based learning method, and model-based learning method. The metric-based learning method, references [26]–[32] focus on classifying well by nearest neighbor classifiers with similarity measurement functions such as Euclidean [30], cosine distance [31], and DeepEMD [32]. For example, DeepEMD [32] splits the picture into multiple tiles and then introduces the earth mover's distance (EMD) as a metric function to obtain more discriminative information with local features. The optimization-based learning method considers whether the machine can learn some training steps by itself. MAML [33] uses a nested optimization learning model to let the machine learn initialization parameters to optimize so that it can quickly adapt to new tasks. MAXL [34] trains two neural networks in the form of dual-gradient meta-learning: a label generation network for predicting auxiliary labels and a multi-task network for training main and auxiliary tasks to improve the generalization ability. Metadock [35] compresses the metamodel by dynamically selecting the kernel, which can be easily deployed on edge devices. The model-based learning methods aim to find the best architecture, in which the model can quickly update parameters. Reference [36] proposed few-shot NAS. The core idea is to divide the super network into several subnetworks to search different areas of the space. Due to the slight increase in the number of super nodes, the accuracy of the few-shot NAS has been greatly improved. In [37], a hierarchical prototype model is introduced, in which each level of the prototype obtains corresponding information from the hierarchical memory, and performs meta-learning on the model through the newly derived hierarchical variational reasoning framework so that the hierarchical memory and the prototype are jointly optimized.

3. Few-shot class-incremental learning

Recently, TOPIC [15] raised a challenging but practical FSCIL problem. To solve this problem, TOPIC proposed a neural gas (NG) network to constrain the feature space topology of knowledge representation and push novel class instances to their respective NG nodes to retain old knowledge and adapt to novel knowledge. Reference [16] proposed a framework to explicitly address the problems of generalized few-shot learning by balancing between learning novel classes, forgetting base classes, and calibration across them in three phases. At the same time, it is based on base-normalized cross-entropy, to overcome the bias learned by the model on the base classes in combination with weight constraints to mitigate the forgetting problem. Reference [17] trained the current task data set by selecting a small number of unimportant parameters, thereby reducing the catastrophic forgetting of old classes. Reference [18] use a non-parametric method based on learning vector quantization in deep embedding space to harmonize old knowledge preservation and novel knowledge adaptation. Reference [19] decoupled the feature extraction module from the classifier and only updated the classifier for each incremental task. And to make the classifier applicable to all classes, they propose a continually evolved classifier (CEC) that employs a graph model to propagate context information between classifiers for adaptation. Reference [20] searched for flat local minima of the objective function during training of the base class, and fine-tuned the model parameters in the flat region of the new task so that the model was adapted to the new classification task. Reference [38] designed an inspatial frequency-aware regularization to enforce SvF constraints on different frequency components, and proposed a spatial combination operation to well balance the slow forgetting of old knowledge and the fast adaptation to new knowledge. Reference [39] proposed the ForwArd compatible training (FACT) strategy by assigning virtual prototypes to compress the embedded representation of base classes, reserve a certain space for new classes, predict possible new tasks, and prepare for the update process. This effectively incorporates new classes with forwarding compatibility while preventing old classes from being forgotten. In reference [40], the method of no data replay is proposed. This method synthesizes data through the generator and applies entropy regularization to encourage more uncertain examples. At the same time, the method uses labels to relabel the generated data, allowing the network to learn only by minimizing the cross-entropy loss, which alleviates the problem of balancing different objectives in traditional knowledge distillation methods. Reference [41] proposes a two-level optimization based on meta-learning to learn how to incrementally learn in the setting of FSCIL. Reference [13] has developed a self-supervised learning (SSL) and knowledge distillation (KD) framework to enhance the feature extraction of the low-capacity backbone network work of ultra-fine-grained FSCIL. Reference [42] proposed the multi-feature space similarity supplement (MFS3) method. This method first trains different feature spaces for different tasks and uses an inter-feature space similarity supplement (IFS3) to focus on the boundary-sensitive sample points to improve the expression ability of each session. At the same time, this method further designed an outer-feature space similarity supplement (OFS3), which can use the new feature space to supplement the basic feature space to re-estimate the sample points.

III. From Old Classes to Novel Classes

1. Problem statement

The FSCIL setting includes a series of labeled training sets $D^1, D^2, \ldots, D^t$, where $D^i = \left\{I^i, C^i\right\}$, t represents tasks, I^i are images corresponding to the i-th task, the corresponding label space of D^i is denoted by C^i, and in the i-th task, only D^i is available. The classes among all tasks are disjoint, i.e., $C^i \cap C^j = \emptyset$ $(i \neq j)$. The first training set, D^1, has N_b classes and enough training samples, termed the old class set, also known as the base class set. In incremental tasks $(t \geq 2)$, there are only a few samples per class, usually described as an N_n-way K-shot training set, where there are N_n classes in the dataset and each class has K samples, termed the novel class set. When task i is trained, the model is evaluated on all encountered classes $C = \bigcup_{i=1}^{t} C^i$.

2. A baseline model for FSCIL

We train the embedding network F' with a classifier C_W [43], [44] by minimizing the cross-entropy loss using the base class D^1, and the embedding network F' is fixed when performing incremental tasks. Specifically, all the training and test samples I are mapped to the embedding space of the embedding network F' to generate embedding vectors: $\boldsymbol{x}' = \mathrm{F}'(I)$. We use the mean vector to calculate a prototype for each class.

$$\boldsymbol{w}'_c = \frac{1}{n_c} \sum_j \left[y_j = c\right] \boldsymbol{x}'_j \tag{1}$$

where c represents the class, and n_c is the number of training samples of class c. If $y_j = c$ is true, then $[y_j = c] = 1$, otherwise it is 0. We can naively form the prototypes of the base class and novel class as $\boldsymbol{W}'_{\text{base}} = \left\{\boldsymbol{w}_1'^{\text{base}}, \boldsymbol{w}_2'^{\text{base}}, \ldots, \boldsymbol{w}_{N_b}'^{\text{base}}\right\} \in \mathbb{R}^{N_b \times d}$ and $\boldsymbol{W}'_{\text{novel}} = \left\{\boldsymbol{w}_1'^{\text{novel}}, \boldsymbol{w}_2'^{\text{novel}}, \ldots, \boldsymbol{w}_{N_n}'^{\text{novel}}\right\} \in \mathbb{R}^{N_n \times d}$. The inference is performed by an NCM classifier with a weight matrix $\boldsymbol{W}' = [\boldsymbol{W}'_{\text{base}}; \boldsymbol{W}'_{\text{novel}}]$. We use Euclidean distance $\operatorname{dist}(\cdot, \cdot)$ to calculate all test samples embeddings $\boldsymbol{X}' = [\boldsymbol{X}'_{\text{base}}; \boldsymbol{X}'_{\text{novel}}]$ and all prototypes $\boldsymbol{W}'$ semantic differences. The classifier is given by

$$f(\boldsymbol{x}, \boldsymbol{W}) = \frac{\mathrm{e}^{-\operatorname{dist}\left(\boldsymbol{x}', \boldsymbol{w}'_c\right)}}{\sum_{\boldsymbol{w}'_j \in \boldsymbol{W}'} \mathrm{e}^{-\operatorname{dist}\left(\boldsymbol{x}', \boldsymbol{w}'_j\right)}} \tag{2}$$

To overcome the problems of embedding bias and classifier bias faced by the baseline model proposed in Section I, we propose the FSCIL-EACA model here. The goal is to learn a robust embedding augmentation network (EAN) that facilitates feature representations for old and new classes, and learn an adaptive incremental classifier (AIC) that dynamically scales new tasks without forgetting old ones.

IV. Methods

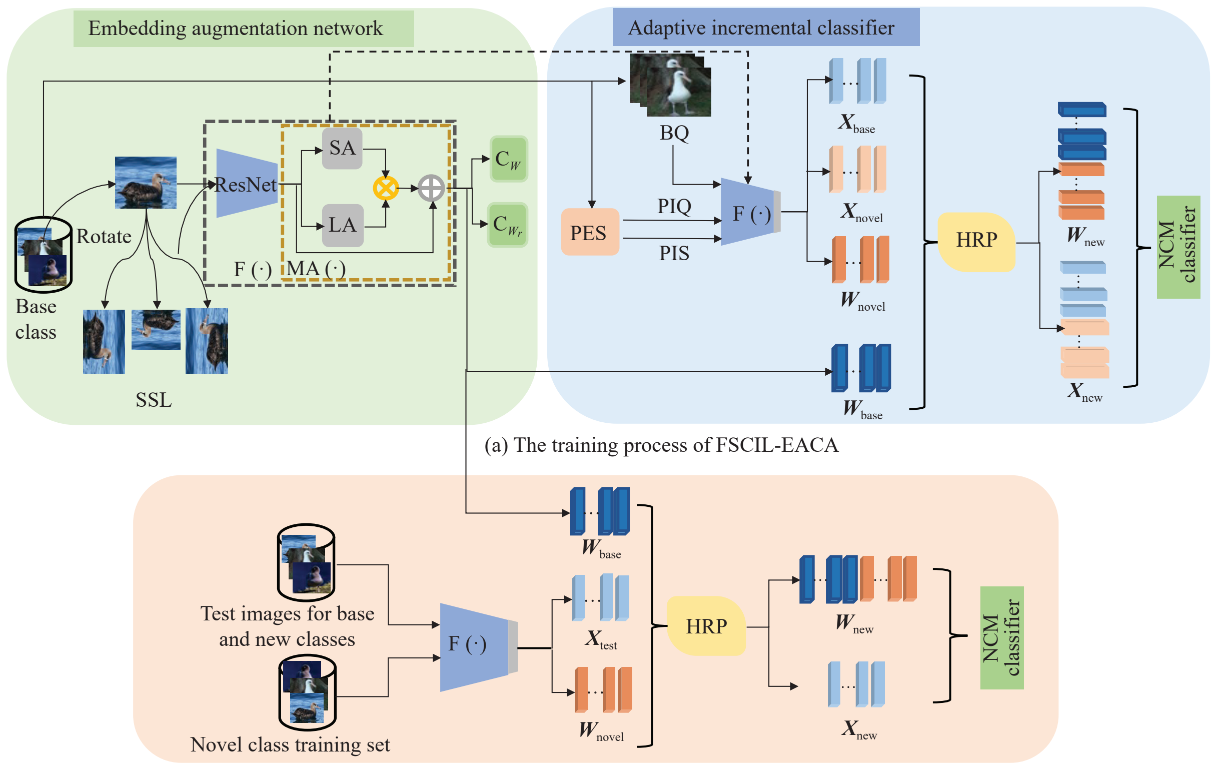
1. Overview

In this section, we introduce FSCIL-EACA in terms of both training and usage. The architecture of our model is shown in Figure 1.

The training process of FSCIL-EACA is shown in Figure 1(a), which mainly includes two stages: 1) Embedding augmentation network (EAN) pre-training stage; 2) Adaptive incremental classifier (AIC) incremental learning stage. Because the data of the base classes and the novel classes are seriously unbalanced, we decouple the two stages. The model only trains EAN in the first base task, and there is a large amount of available data in the base task. In the novel task, we fixed the EAN, and adjusted the AIC according to the novel and old classes, thus minimizing the over-fitting and catastrophic forgetting problems at the task level. Details are given next.

EAN pre-training stage In the base task, we introduce the self-supervised learning (SSL) and modulation attention (MA) modules and train the EAN $\mathrm{F}(\cdot)$ network through the base class dataset.

AIC incremental learning stage In order to realize the adaptive adjustment of novel and old classes in incremental tasks, we propose the hybrid relation projection (HRP) module, and we train it through meta-learning to quickly adapt to novel classes. Specifically, first, we generate pseudo-incremental data through the pseudo-incremental episode selection (PES) module, including the



(a) The training process of FSCIL-EACA

(b) The use process of FSCIL-EACA

Figure 1 Overall network architecture.

pseudo-incremental support set (PIS) and pseudo-incremental query set (PIQ). Since we hope that HRP can classify novel and old classes, we save prototypes (usually the mean of the feature embeddings of all samples in the class) $\boldsymbol{W}_{\mathrm{base}}$ for all base classes, and also extract the base query set (BQ) from the base class to jointly train the HRP module. Then, input PIS, PIQ, and BQ into the EAN module to get $\boldsymbol{W}_{\mathrm{novel}}$, $\boldsymbol{X}_{\mathrm{novel}}$, and $\boldsymbol{X}_{\mathrm{base}}$, respectively. Finally, we project $\boldsymbol{W}_{\mathrm{base}}$ and $\boldsymbol{W}_{\mathrm{novel}}$ to the prototype self-projection (PSP) module of HRP training update to get updated prototypes $\boldsymbol{W}_{\mathrm{new}}$. And project $\boldsymbol{W}_{\mathrm{base}}$, $\boldsymbol{W}_{\mathrm{novel}}$, $\boldsymbol{X}_{\mathrm{novel}}$, and $\boldsymbol{X}_{\mathrm{base}}$ to the query set cross-projection (QCP) module of HRP training update to get $\boldsymbol{X}_{\mathrm{new}}$ and classify by NCM classifier. The parameters are updated by cross-entropy loss. Equation (2) is rewritten as

$$f(\boldsymbol{x},\boldsymbol{W})=\frac{\mathrm{e}^{-\mathrm{dist}(\boldsymbol{x}_{\mathrm{new}},\boldsymbol{w}^{c}_{\mathrm{new}})}}{\sum_{\boldsymbol{w}^{j}_{\mathrm{new}}\in\boldsymbol{W}_{\mathrm{new}}}\mathrm{e}^{-\mathrm{dist}\left(\boldsymbol{x}_{\mathrm{new}},\boldsymbol{w}_{\mathrm{new}}{}^{j}\right)}} \tag{3}$$

The use process of FSCIL-EACA is shown in Figure 1(b). After FSCIL-EACA training, the model has the ability to continuously learn. We freeze EAN and HRP modules and deploy them to real incremental tasks. When a novel class appears, we can directly get the prototype representation $\boldsymbol{W}_{\mathrm{novel}}$ of the novel class through EAN, and the new image will no longer undergo SSL-based rotation change and prediction process. Finally, $\boldsymbol{W}_{\mathrm{novel}}$ and base class prototypes $\boldsymbol{W}_{\mathrm{base}}$ input PSP module to get updated $\boldsymbol{W}_{\mathrm{new}}$. When there is an image to be classified, we input the embedded representation $\boldsymbol{X}_{\mathrm{test}}$ and prototype $\boldsymbol{W}_{\mathrm{novel}}$ of the image into QCP together with the base class prototypes $\boldsymbol{W}_{\mathrm{base}}$ to obtain the updated $\boldsymbol{X}_{\mathrm{new}}$. Finally, the classification is realized by the NCM classifier. At the same time, we merged $\boldsymbol{W}_{\mathrm{novel}}$ into $\boldsymbol{W}_{\mathrm{base}}$ as the base class prototypes for the next novel task, i.e., $\boldsymbol{W}_{\mathrm{base}}=[\boldsymbol{W}_{\mathrm{base}};\ \boldsymbol{W}_{\mathrm{novel}}]$.

2. Embedding augmentation network (EAN)

Self-supervised learning (SSL) The standard classification network is directly trained from a large labeled dataset to learn the weight vectors of different classes and then classify test samples. However, in our few-shot class-incremental task, the embedding representations of the novel class samples are calculated by the embedding function trained in the base classes. In other words, the embedded knowledge learned on the base data set needs to be well transferred to the novel classes for classification prediction. Therefore, we believe that the embedding function needs to have good generalization and transferability and generate robust feature embeddings for invisible classes.

SSL has been proven to use auxiliary tasks to mine its supervision information from large-scale unsupervised data to improve the generalization ability and robustness of the model to learn valuable representations for downstream tasks [45]. Recently, the SSL method based on rotation prediction has achieved great success in related scenarios such as incremental learning [12], few-shot incremental learning [17], and class imbalance classification [46], so we use rotation prediction as our auxiliary task. In this work, we rotate the images I by r degrees and obtain $\boldsymbol{x}^r$ by the embedding network F, where $r \in \{0°, 90°, 180°, 270°\}$. We add an additional 4-way classifier C_{W_r} to predict one of 4 classes in r. The loss is given by

$$L\left(D^1\right) = \left[L_{\mathrm{CE}}\left(\mathrm{C}_{W_r}\left(\boldsymbol{x}^r\right), r\right) + L_{\mathrm{CE}}\left(\mathrm{C}_W\left(\boldsymbol{x}^r\right), y\right)\right]/2 \tag{4}$$

where L_{CE} is the standard cross-entropy loss.

Modulated attention (MA) Although SSL helps generalize features between old and new classes, it is also important to extract discriminative features. Therefore, we introduce modulation attention [47] to obtain more representative feature embeddings and maintain the differences between classes.

The modulation attention $\mathrm{MA}(\boldsymbol{x})$ consists of two parts, self-attention $\mathrm{SA}(\boldsymbol{x})$ and location attention $\mathrm{LA}(\boldsymbol{x})$. First, the interaction information between any two positions in the feature map is calculated by self-attention [48] to capture the long-range dependency, which makes each pixel of the feature map contain contextual information, and the $\mathrm{SA}(\boldsymbol{x})$ map is obtained. Then, $\mathrm{LA}(\boldsymbol{x})$ is applied on the basis of $\mathrm{SA}(\boldsymbol{x})$ to select the information of some positions with the most discriminative ability and encourage different classes to use context information of different strengths, which helps maintain the distinction between old and new classes. In the article, $\mathrm{LA}(\boldsymbol{x})$ is mainly implemented by a fully connected layer + softmax function. First, the original feature map is flattened, then the weight information map of different feature space positions is learned through the fully connected layer and the softmax function, and finally, the weight information map is restored to the original feature map size to obtain $\mathrm{LA}(\boldsymbol{x})$ map. The final feature map becomes

$$\mathrm{MA}(\boldsymbol{x}) = \mathrm{LA}(\boldsymbol{x}) \otimes \mathrm{SA}(\boldsymbol{x}) + \boldsymbol{x} \tag{5}$$

where $+\boldsymbol{x}$ represents a residual connection, the contextual information is added back (by skipping the connection) to the original feature map for enhancement. We apply MA to the last layer of the feature embedding network.

3. Hybrid relation projection (HRP)

Transformers [49], [50] can learn contextual relations between all prototypes and query set embeddings without considering their ordering, which is suitable for modeling complex interactions between prototypes and query set embeddings. So, we developed transformer-based HRP modules, including prototype self-projection (PSP) and query set cross-projection (QCP), as shown in Figure 2.

Prototype self-projection (PSP) In order to establish a global dependency between the previous and current task prototypes, the adaptive adjustment of the prototype is realized to make the prototypes more distinguishable in the current task space. We train a set-to-set function $\mathrm{T}(\cdot)$ based on Transformers, which transforms a set of original prototypes $\boldsymbol{W}$ into a set of updated prototypes $\boldsymbol{W}_{\mathrm{new}}$, where $\boldsymbol{W}_{\mathrm{new}} = T(\boldsymbol{W})$. The inputs of $T(\cdot)$ adopt the triple form of $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ share the same input source $\boldsymbol{W}$, PSP can be expressed as

$$\boldsymbol{Q} = \boldsymbol{W}_Q\boldsymbol{W} \qquad \boldsymbol{K} = \boldsymbol{W}_K\boldsymbol{W} \qquad \boldsymbol{V} = \boldsymbol{W}_V\boldsymbol{W} \tag{6}$$

$$\tilde{\boldsymbol{V}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{m}}\right)\boldsymbol{V} \tag{7}$$

$$\boldsymbol{W}_{\mathrm{new}} = \mathrm{LayerNorm}\left(\mathrm{Dropout}\left(\boldsymbol{W}_{\mathrm{FC}}\tilde{\boldsymbol{V}} + \boldsymbol{W}\right)\right) \tag{8}$$

where $\boldsymbol{W} = [\boldsymbol{W}_{\mathrm{base}}; \boldsymbol{W}_{\mathrm{novel}}]$ indicates that the old class prototypes and the novel class prototypes are merged, and updated simultaneously through PSP. The matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{m\times hm}$ are the learnable transform matrices, which project the original prototypes into the shared metric space, h is the number of heads, $\boldsymbol{W}_{FC} \in \mathbb{R}^{hm\times m}$ is a trainable weight matrix, and m is the embedding dimension. $\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{m}}\right)\boldsymbol{V}$ is the relation matrix between prototypes. We use the relation matrix as the weight coefficients to aggregate the information from all the prototypes in $\boldsymbol{V}$ to obtain $\tilde{\boldsymbol{V}}$. Finally, the original prototypes are merged to obtain updated prototypes $\boldsymbol{W}_{\mathrm{new}}$.

Query set cross-projection (QCP) After the classifier prototypes are updated, it will inevitably lead to a mismatch between the new and old class feature representations and the classifier, so the feature representation needs to be adjusted to better adapt to the current classification task. Because the query set samples need to be classified according to the distance from the prototypes, we establish a cross-projection process between the query set embeddings $\boldsymbol{X}$ and the prototypes $\boldsymbol{W}$ to obtain the updated embeddings $\boldsymbol{X}_{\mathrm{new}}$. QCP and PSP use the same calculation form, i.e., $\boldsymbol{X}_{\mathrm{new}} = T(\boldsymbol{X}, \boldsymbol{W})$. The formula is as follows:

$$\boldsymbol{Q} = \boldsymbol{W}_Q\boldsymbol{X}, \quad \boldsymbol{K} = \boldsymbol{W}_K\boldsymbol{W}, \quad \boldsymbol{V} = \boldsymbol{W}_V\boldsymbol{W} \tag{9}$$

$$\tilde{\boldsymbol{X}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{m}}\right)\boldsymbol{V} \tag{10}$$

$$\boldsymbol{X}_{\mathrm{new}} = \mathrm{LayerNorm}\left(\mathrm{Dropout}\left(\boldsymbol{W}_{\mathrm{FC}}\tilde{\boldsymbol{X}} + \boldsymbol{X}\right)\right) \tag{11}$$

4. Pseudo-incremental episode selection (PES)

In the FSCIL scenario, it is very important to train HRP to have the ability of continuously learning. We propose a method based on meta-learning to conduct multi-stage training for HRP to achieve this goal. Meta-learning requires a large number of support sets and query sets to learn the common characteristics of cross-tasks. However, in FSCIL, only data from a single task is available, and the amount of data in incremental tasks is

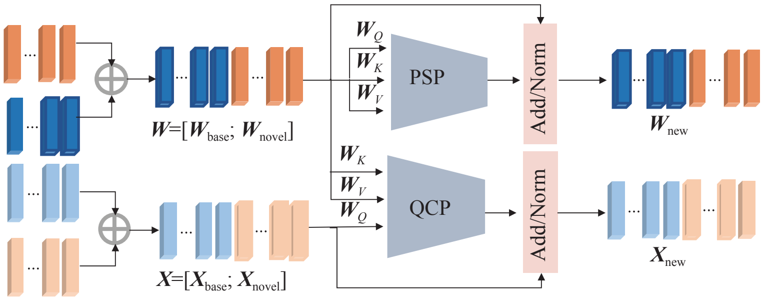

W_Q
W_K
W_V
PSP
Add/Norm
W=[W_{base}; W_{novel}]
W_{new}
W_K
W_V
W_Q
QCP
Add/Norm
X=[X_{base}; X_{novel}]
X_{new}

Figure 2 Hybrid relation projection module.

always limited, which cannot be used to support meta-learning training methods. Fortunately, in the FSCIL task, base data is sufficient, so how to make full use of base data for meta-learning training plays a key role in HRP's continuous learning ability. In [14], in order to make the embedded model see more categories during training and improve the portability of the model, a large number of novel classes were generated based on the method of mixup [51]. Inspired by [14], we design a pseudo-incremental episode selection (PES) algorithm to construct pseudo-incremental tasks based on base classes to simulate incremental scenes, as shown in Figure 3. PES randomly draws two samples, I_a and I_b, from two different base classes a and b to generate a representative pseudo-incremental class (PIC) sample I_{ab}.

$$I_{ab} = \lambda I_a + (1 - \lambda) I_b \tag{12}$$

where λ is a random number of interpolations coefficient, based on N_b base classes, we can use the above method to generate $N_b\left(N_b - 1\right)/2$ pseudo-incremental new classes. Similar to the episodic meta-learning strategy, we use pseudo-incremental datasets to generate a pseudoincremental support set (PIS) in the form of an N_n-way K-shot in each iteration. The corresponding pseudo-incremental query set (PIQ) is sampled from N_n classes.

V. Experiments

1. Dataset

In the experiment, we selected two classic datasets, CUB-200 and MiniImageNet, that are widely used to evaluate the performance of FSCIL, as well as two fine-grained medical datasets, HyperKvasir and SKIN-7.

The CUB-200 dataset contains approximately 6,000 training images and 6,000 test images from 200 classes. Each image size is 224 × 224. We split the 200 classes into 100 base classes and 100 incremental classes. For each incremental session training set of CUB-200, we use the 10-way 5-shot setting. Hence, for the CUB-200 dataset, we have 1 base session and 10 incremental sessions (11 sessions in total).

The MiniImageNet dataset contains 60,000 images from 100 classes, which are selected from the ImageNet dataset. Each class has 500 training images and 100 test images, with a size of 84 × 84. We split the 100 classes into 60 base classes and 40 incremental classes. For each incremental session, we use the 5-way 5-shot setting. Therefore, for the MiniImageNet dataset, we have 1 base session and 8 incremental sessions (a total of 9 sessions).

The HyperKvasir dataset [52] is one of the largest publicly available gastrointestinal endoscopy datasets under CC BY 4.0 (Creative Commons Attribution 4.0 International). The dataset includes labeled images, segmented images, unlabeled images, and labeled videos, and we choose labeled images among them for experiments. The dataset contains a total of 10,662 labeled images, and 23 categories, including BBPS 2-3 (1148), Polyps (1028), Cecum (1009), Dyed lifted polyps (1002), Pylorus (999), Dyed resection margins (989), Z-line (932), Retroflex stomach (764), BBPS 0-1 (646), Ulcerative colitis grade-2 (443), Esophagitis grade A (403), Retroflex rectum (391), Esophagitis grade B-D (260), Ulcerative colitis grade-1 (201), Ulcerative colitis grade-3 (133), Impacted stool (131), Barrett's short segments (53), Barretts (41), Ulcerative colitis grade-0-1 (35), Ulcerative colitis grade-2-3 (28), Ulcerative colitis grade-1-2 (11), Ileum (9), Hemorrhoids (6). Among them, the number of samples for the latter three diseases is too small, so we selected the first 20 diseases to carry out the experiment. We selected the top 5 disease categories with the largest amount of data as the base classes, and the remaining 15 as the incremental classes. In the HyperKvasir dataset, each incremental task had 3 new classes, with 10 images per category, which means the new class dataset was set at 3-way 10-shot (a total of 6 sessions).

The SKIN-7 dataset [53] contains the 7 most important skin disease classes in the realm of pigmented lesions collected from different age groups, different regions, and different methods. There are 10,015 dermatoscopic images in total. The images are labeled by expert pathologists as Melanocytic nevi (NV, 6705), Melanoma (MEL, 1113), Benign keratosis (BKL, 1099), Basal cell carcinoma (BCC, 514), Actinic keratoses and Intraepithelial carcinoma (AKIEC, 327), Vascular lesion (VASC, 142) and Dermatofibroma (DF, 115). We chose NV, MEL, and BKL as the base classes and the remaining 4 classes as the increment classes. For each incremental session of SKIN-7, we use settings of 2-way 10-shot for experiments. Therefore, for the SKIN-7, we have 1 base session and 2 incremental sessions (a total of 3 sessions).

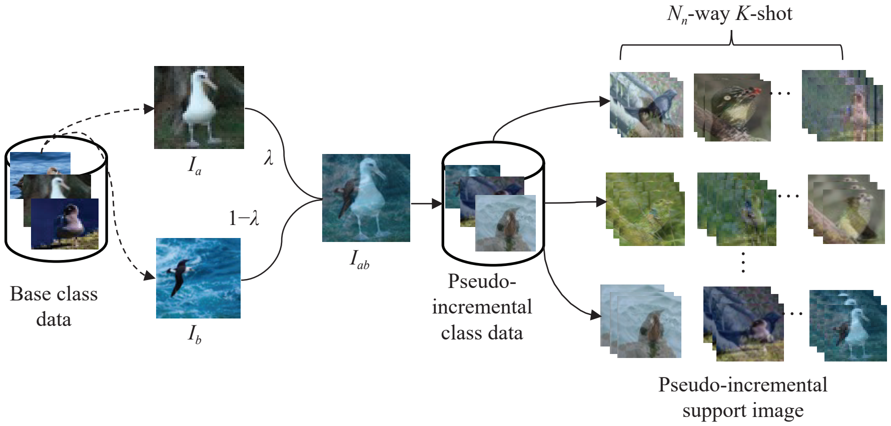


Figure 3 Pseudo-incremental Episode Selection module.

2. Implementation details

In the first stage, we use the ResNet-18 architecture as the backbone network for both MiniImageNet and CUB-200 datasets, pre-train 100 epochs on the base task with the SGD optimizer, and use the same initial learning rate as CEC, which is 0.1. After 30 and 40 epochs, we reduced the learning rate to 0.01 and 0.001, respectively. We found that when the initial learning rate was adjusted to 0.01, better results were achieved on the two natural image datasets, so we carried out experiments with different learning rates. For HyperKvasir and SKIN-7, we use the ResNet-34 architecture as the backbone network with a learning rate of 0.01, and other settings are the same as the natural image experiments.

In the second stage, for MiniImageNet and CUB-200, the support set selection method is 15-way 1-shot. For HyperKvasir, it's 5-way 1-shot. For SKIN-7, it's 2-way 5-shot. We fix E to 10, use an initial learning rate of 0.0002 on all datasets, and train on HRP for 100 epochs.

3. Comparison with state-of-the-art methods

To better evaluate the overall performance of our model, in addition to the baseline introduced, we compare FSCIL-EACA with state-of-the-art methods for FSCIL (SSFE-Net [13], TOPIC [15], FSLL [17], CEC [19], MgSvF [38], Data-free replay [40], MetaFSCIL [41], and MFS3 [42]) and some classic methods of CIL (iCaRL [5], EEIL [7], NCM [11], PASS [12], and IL2A [14]). At the same time, we also compared the Ft-CNN and Joint-CNN models. The Ft-CNN model only involves fine-tuning the model on incremental tasks; the Joint-CNN model involves co-training on-base task data and incremental task data. We first conducted an experimental comparison on the CUB-200 and MiniImageNet datasets to prove the superiority of our model. Then we validated our method on the HyperKvasir and SKIN-7 datasets.

Results and discussion on CUB-200 The results of CUB-200 are shown in Table 1. All models in the experiments were pre-trained with ImageNet. According to the experimental results, we have the following observations:

1) The baseline model has significant superiority. Our model FSCIL-EACA proposes EAN and AIC modules on the basis of the baseline model, and the average accuracy (Avg) is improved by 4.73%.

2) FSCIL-EACA significantly outperforms Ft-CNN and Joint-CNN models. The Avg of FSCIL-EACA on the CUB-200 dataset is 36.21% and 16.50% higher than that of Ft-CNN and Joint-CNN, respectively. This shows that fine-tuning on new data or retraining on all old and new data is not the best option when there are only a few new data.

3) Our model is better than typical incremental learning methods like iCaRL, EEIL, and NCM. This proves that our method can effectively solve the catastrophic forgetting problem in FSCIL without storing old data samples, and only a few samples of new classes are needed to adapt to new tasks, whereas incremental learning methods require sufficient samples for each new class to obtain satisfactory performance.

4) Compared with typical FSCIL methods such as TOPIC, FSLL, CEC, etc., the performance of FSCIL-EACA has improved in almost all sessions. Compared with the most similar CEC model, the Avg has increased by 3.45%. It shows that FSCIL-EACA performs better through the embedding augmentation and classifier adaptation strategy.

5) Better results are achieved when the initial learning rate is 0.01. The * in the following experimental results represents that we have adjusted the learning rate. We denote the experimental results as FSCIL-EACA*. Compared with FSCIL-EACA, the Avg of FSCIL-EACA* has increased by 1.68%. Compared to models with the same initial learning rate of 0.01 (FSLL*, FSLL+SSL*, IDLVQ-C*), our model has absolute advantages.

Results and discussion on MiniImageNet Table 2 shows the results of the experiments. From the related results, our model also achieves excellent results on MiniImageNet and can draw the same conclusions as CUB-200. It is worth noting that none of the current state-of-the-art methods are pre-trained, so only the models identified by # are pre-trained in our experiments (i.e., FSCIL-EACA# and FSCIL-EACA*#). It can be seen from the results that further improvement has been achieved

Table 1 Results on CUB-200 using the ResNet-18 architecture on 10-way 5-shot FSCIL setting

Tasks												
Model	1	2	3	4	5	6	7	8	9	10	11	Avg
Baseline	75.92	72.02	67.37	62.68	61.15	57.67	56.51	53.58	52.40	51.39	49.89	60.05
Ft-CNN	68.68	44.81	32.26	25.83	25.62	25.22	20.84	16.77	18.82	18.25	17.18	28.57
Joint-CNN	68.68	62.43	57.23	52.80	49.50	46.10	42.80	40.10	38.70	37.10	35.60	48.28
iCaRL [5]	68.68	52.65	48.61	44.16	36.62	29.52	27.83	26.26	24.01	23.89	21.16	36.67
EEIL [7]	68.68	53.63	47.91	44.20	36.30	27.46	25.93	24.70	23.95	24.13	22.11	36.27
NCM [11]	68.68	57.12	44.21	28.78	26.71	25.66	24.62	21.52	20.12	20.06	19.87	32.49
TOPIC [15]	68.68	62.49	54.81	49.99	45.25	41.40	38.35	35.36	32.22	28.31	26.28	43.92
FSLL [17]	68.72	65.67	62.33	58.10	55.44	52.66	51.17	50.27	48.31	47.25	45.55	55.04
CEC [19]	75.85	71.94	68.50	63.50	62.43	58.27	57.73	55.81	54.83	53.52	52.28	61.33
Data-Free Replay [40]	75.90	72.14	68.64	63.76	62.58	59.11	57.82	55.89	54.92	53.58	52.39	61.52
MetaFSCIL [41]	75.90	72.41	68.78	64.78	62.96	59.99	58.30	56.85	54.78	53.82	52.64	61.92
MgSvF [38]	72.29	70.53	67.00	64.92	62.67	61.89	59.63	**59.15**	**57.73**	55.92	54.33	62.37
MFS3 [42]	75.63	72.51	69.65	65.29	63.13	60.38	58.99	57.41	55.55	54.95	53.47	62.45
SSFE-Net [13]	76.38	72.11	68.82	64.77	63.59	60.56	59.84	58.93	57.33	56.23	54.28	62.99
FSCIL-EACA	**79.04**	**75.19**	**71.61**	**66.59**	**66.04**	**62.71**	**60.57**	59.05	57.66	**57.46**	**56.61**	**64.78**
FSLL* [17]	72.77	69.33	65.51	62.66	61.10	58.65	57.78	57.26	55.59	55.39	54.21	60.93
FSLL+SSL* [17]	75.63	71.81	68.16	64.32	62.61	60.10	58.82	58.70	56.45	56.41	55.82	62.62
IDLVQ-C* [18]	77.37	74.72	70.28	67.13	65.34	63.52	62.10	61.54	59.04	58.68	57.81	65.23
FSCIL-EACA*	**79.25**	**76.03**	**72.41**	**67.81**	**67.76**	**64.84**	**64.07**	**62.13**	**61.13**	**59.98**	**59.32**	**66.79**

Table 2 Results on MiniImageNet using the ResNet-18 architecture on 5-way 5-shot FSCIL setting

Tasks										
Model	1	2	3	4	5	6	7	8	9	Avg
Baseline	70.67	66.09	61.94	58.59	55.61	52.73	50.12	48.32	46.99	56.78
Baseline#	73.90	68.91	64.73	61.36	58.44	55.31	52.66	50.74	48.98	59.45
Ft-CNN	61.31	27.22	16.37	6.08	2.54	1.56	1.93	2.60	1.40	13.45
Joint-CNN	61.31	56.60	52.60	49.00	46.00	43.30	40.90	38.70	36.80	47.25
iCaRL [5]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.8	17.21	33.29
EEIL [7]	61.31	46.58	44.00	37.29	33.14	27.12	24.10	21.57	19.58	34.97
NCM [11]	61.31	47.8	39.31	31.91	25.68	21.35	18.67	17.24	14.17	30.83
TOPIC [15]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	39.64
FSLL [17]	61.32	58.43	54.53	50.68	47.79	45.58	43.36	40.93	39.49	49.12
CEC [19]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	57.75
SSFE-Net [13]	72.06	66.17	62.25	59.74	56.36	53.85	51.96	49.55	47.73	57.74
Data-Free Replay [40]	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	58.02
MetaFSCIL [41]	72.04	67.94	63.77	60.29	57.58	55.16	52.9	50.79	49.19	58.85
MFS3 [42]	73.65	68.91	64.60	61.48	58.68	55.55	53.33	51.69	50.26	59.79
FSCIL-EACA	**76.23**	**71.11**	**66.31**	**63.39**	**60.60**	**57.13**	**54.37**	**52.19**	**50.61**	**61.33**
FSCIL-EACA&	**77.45**	**72.15**	**67.9**	**64.41**	**61.40**	**58.32**	**55.44**	**53.45**	**51.70**	**62.47**
FSCIL-EACA#	**76.18**	**71.32**	**67.40**	**64.17**	**61.48**	**58.25**	**55.46**	**53.25**	**51.83**	**62.15**
FSLL* [17]	66.48	61.75	58.16	54.16	51.10	48.53	46.54	44.20	42.28	52.58
FSLL+SSL* [17]	68.85	63.14	59.24	55.23	52.24	49.65	47.74	45.23	43.92	53.92
IDLVQ-C* [18]	64.77	59.87	55.93	52.62	49.88	47.55	44.83	43.14	41.84	51.16
FSCIL-EACA*#	**80.50**	**75.89**	**71.46**	**68.00**	**65.23**	**61.88**	**58.71**	**56.59**	**55.14**	**65.93**

compared to FSCIL-EACA. We added an additional experiment on MiniImageNet, replacing the SSL method with the Mixup data augmentation method. We perform Mixup data augmentation on the 60 base classes to obtain $60 \times (60-1)/2$ enhanced classes. In this way, the original 60-class problem in the base task is expanded to $60 + 60 \times (60-1)/2$-class questions to train the embedding augmentation network to learn transferable and diverse embedding representations. For the PES module, we use three types of base image merging to obtain pseudo-incremental classes, i.e., $I_{abc} = \lambda I_a + \beta I_b + (1-\lambda-\beta) I_c$, where λ and β are random numbers of interpolation coefficients. In the same way, we deploy the model to real incremental tasks, and do not perform the Mixup data augmentation process in the new data classification process. The experimental results are expressed as FSCIL-EACA&, and the Avg is 1.14% higher than that of FSCIL-EACA. This method only works with MiniImageNet. There is no improvement when we adopt this method in CUB-200. This may be due to the large number of CUB-200 classes. This method expands the basic tasks to $100 + 100 \times (100-1)/2$-class, and each base class has only 60 samples, which is not conducive to feature representation.

Results and discussion on HyperKvasir The experimental results are shown in Table 3. Compared with Baseline, Joint-CNN, and CEC, FSCIL-EACA has achieved the same effect on fine-grained medical image datasets as in natural image datasets, showing the superiority of the FSCIL-EACA model. Compared with the model IL2A and PASS network, which also uses the enhanced algorithm in the embedded network, FSCIL-EACA has achieved better results, which proves the effectiveness of the embedded enhanced network proposed in this paper. In subsequent tasks, FSCIL-EACA only needs a small number of new class samples to adapt to new tasks, while IL2A and PASS need a large amount of new task data to achieve satisfactory results.

Table 3 Results on HyperKvasir using the ResNet-34 architecture on 3-way 10-shot FSCIL setting

Model	Tasks						
	1	2	3	4	5	6	Avg
Acc							
Baseline	98.82	88.98	79.94	74.21	67.59	67.51	79.51
Joint-CNN	98.82	87.89	78.33	73.63	68.68	68.06	79.24
IL2A [14]	97.78	80.23	63.41	55.92	46.25	36.72	63.39
PASS [12]	98.07	82.26	62.77	43.88	31.58	18.15	56.19
CEC [18]	99.09	89.01	80.02	73.99	67.86	67.91	79.65
FSCIL-EACA	**99.27**	**90.71**	**82.82**	**77.53**	**72.71**	**71.04**	**82.35**
F1							
Baseline	98.75	89.00	80.08	75.17	70.61	69.06	80.45
Joint-CNN	98.75	88.18	78.69	75.32	71.75	70.15	80.47
IL2A [14]	97.35	81.22	64.04	57.86	49.45	38.62	64.76
PASS [12]	97.77	82.56	63.24	46.35	35.21	20.75	57.65
CEC [19]	99.04	88.94	80.24	74.88	70.89	69.35	80.56
FSCIL-EACA	**99.19**	**90.71**	**83.33**	**78.99**	**75.21**	**72.59**	**83.34**
AUC							
Baseline	99.22	93.72	89.04	86.63	84.39	83.72	89.45
Joint-CNN	99.22	93.24	88.28	86.71	84.99	84.29	89.46
IL2A [14]	98.61	88.70	79.87	76.27	71.45	66.69	80.27
PASS [12]	98.79	89.87	79.52	69.78	63.65	56.92	76.42
CEC [19]	99.40	93.68	89.13	86.47	84.54	83.87	89.52
FSCIL-EACA	**99.54**	**94.70**	**90.83**	**88.69**	**86.83**	**85.57**	**91.03**

Results and discussion on SKIN-7 The results in Table 4 show that our model achieves better and more stable results, with consistent increases in all three evaluation metrics of Accuracy (Acc), F1, and receiver operating characteristic (ROC) curve (AUC). This shows our model's excellent ability to deal with real-world medical data problems.

4. Ablation experiments

The proposed method comprises three components: SSL, MA, and AIC. Here we conducted ablation experiments on the CUB-200 and MiniImageNet datasets and further discussed the impact of each component.

For CUB-200, we conducted four different experiments with the following experimental settings: the mod-

Table 4 Results on SKIN-7 using the ResNet-34 architecture on 2-way 10-shot FSCIL setting

Model	Tasks								
	1			2			3		
	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Basline	91.72	91.76	93.82	80.33	81.38	88.36	75.66	75.81	85.89
CEC [19]	92.01	92.04	94.03	79.04	79.53	87.21	74.44	74.87	84.88
FSCIL-EACA	**92.53**	**92.65**	**94.49**	**80.53**	**81.87**	**88.67**	**78.24**	**78.42**	**87.41**

el contains only the AIC module, only the SSL module with feature generalization, contains both the SSL and MA modules (i.e., EAN), contains both the EAN and AIC modules (i.e., EAN+AIC, EAN+AIC*). Table 5 illustrates the experimental results for these four variants. When the AIC module is added to the baseline, the classification performance on each task is significantly improved, and the average accuracy rate is increased by 2.16%. This suggests that AIC models global task correlations, facilitating efficient integration of unified classifiers. And use the global task information to accurately update the embedding representation of the test sample to make it more suitable for the current classification task. We observed some incremental precision when adding the SSL module to baseline. When using SSL and MA modules (i.e., EAN) for feature embedding enhancement, more incremental accuracy is produced, and the average accuracy is 3.62% higher than baseline. The effectiveness of EAN shows that a well-trained embedding augmentation network is beneficial to improving the stability of the model and is of great value for extracting the embedding features of the base class and new class. When the EAN and AIC modules are introduced simultaneously, the performance on each task is further improved, and the average accuracy of EAN+AIC is 4.73% higher than that of the baseline. It is shown that the EAN and AIC modules promote each other, and the combination of the two significantly improves the classification performance of FSCIL.

For the ablation experiments on MiniImageNet and HyperKvasir datasets as shown in Tables 6 and 7, we can get some conclusions similar to the CUB-200 experiments.

5. Visualization

Visual class activation mapping (CAM) is a popular tool for visualizing discriminative regions. To visualize the MA module effect, we adapted the post-hoc visual explanation method (Score-CAM) [54]. The results in Figure 4 show that the MA module can capture more discriminative features from the collaborative context, clearly focusing on the target object. Therefore, the visualization intuitively demonstrates the effectiveness of the MA module.

We use t-SNE to visualize the prototype and test set feature embeddings for each class on the CUB-200 dataset. As shown in Figure 5, the prototypes are denoted by “▲”, the test set embeddings are denoted by “●”, and the dots with different colors represent different data classes. In this study, five classes are randomly selected as base classes (30, 47, 58, 78, 88), and five additional classes are added as incremental classes (108, 128, 138,

Table 5 Ablation experiment results on CUB-200

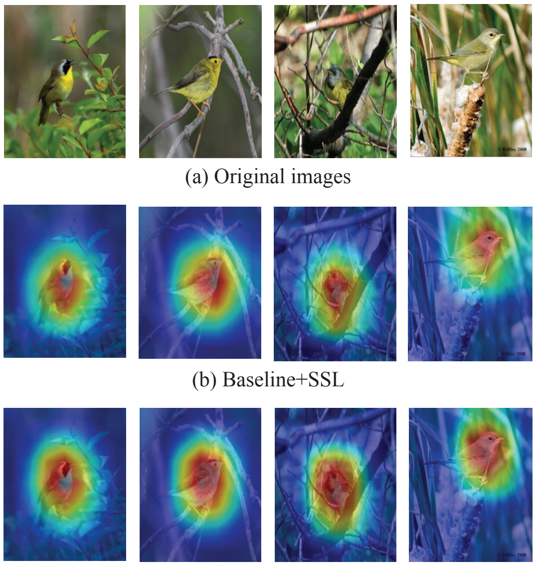
Model	Tasks											
	1	2	3	4	5	6	7	8	9	10	11	Avg
Baseline	75.92	72.02	67.37	62.68	61.15	57.67	56.51	53.58	52.40	51.39	49.89	60.05
+AIC	77.66	73.39	69.47	64.31	63.33	59.79	58.88	56.16	54.70	54.05	52.52	62.21
+SSL	78.12	73.50	68.76	63.86	62.70	58.59	56.74	56.14	52.94	52.98	51.59	61.45
+EAN	78.67	74.66	70.62	65.62	64.68	61.15	59.74	58.72	56.36	55.61	54.56	63.67
+EAN+AIC	79.04	75.19	71.61	66.59	66.04	62.71	60.57	59.05	57.66	57.46	56.61	64.78
+EAN+AIC*	79.25	76.03	72.41	67.81	67.76	64.84	64.07	62.13	61.13	59.98	59.32	66.79

Table 6 Ablation experiment results on MiniImageNet

Tasks										
Model	1	2	3	4	5	6	7	8	9	Avg
Baseline#	73.90	68.91	64.73	61.36	58.44	55.31	52.66	50.74	48.98	59.45
+AIC#	75.43	70.29	65.97	62.44	59.49	56.26	53.58	51.43	49.59	60.50
+EAN#	75.23	70.06	66.06	62.65	59.50	56.72	53.93	52.02	50.65	60.76
+EAN+AIC#	76.18	71.32	67.40	64.17	61.48	58.25	55.46	53.25	51.83	62.15
+EAN+AIC*#	80.50	75.89	71.46	68.00	65.23	61.88	58.71	56.59	55.14	65.93

Table 7 Ablation experiment results on HyperKvasir

Model	Tasks						
	1	2	3	4	5	6	Avg
Baseline	98.82	88.98	79.94	74.21	67.59	67.51	79.51
+MA	98.84	88.99	80.97	75.89	69.91	68.76	80.56
+EAN	99.09	90.90	82.25	76.24	71.12	68.98	81.43
+EAN +AIC	99.27	90.71	82.82	77.53	72.71	71.04	82.35

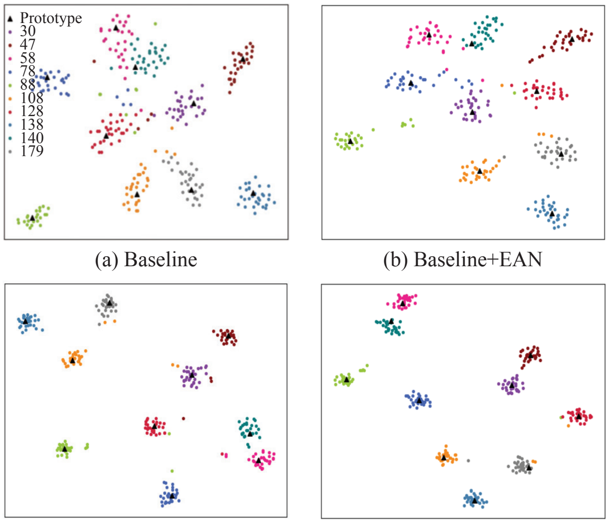


(a) Original images

(b) Baseline+SSL

(c) Baseline+SSL+MA

Figure 4 The visualized activation maps of MA on CUB-200.



(a) Baseline

(b) Baseline+EAN

(c) Baseline+AIC

(d) Baseline+EAN+AIC

Figure 5 Prototype and test set feature embedding visualization.

140, 179). We show results for two configurations: 1) with and without EAN and 2) with and without AIC. In Figure 5(a), we show the baseline model before the incremental update, and Figure 5(b) shows the baseline+EAN model before the incremental update. We observe that EAN reduces the overlap between the base class and the new class (such as the base class 58 and the new class 140). And make the new classes more compact in the feature space (such as new classes 128 and 140). Therefore, EAN can learn more general and transferable features for downstream tasks, resulting in robust embedding representations. Then we add the AIC module to incrementally update the prototypes and test set feature embeddings, as visualized in Figure 5(c) and Figure 5(d). We can infer that the AIC module further improves the distribution of old and new classes, contributing to the higher density of samples of the same class in the feature space. And the prototypes can be adaptively updated to calibrate the decision boundary between old and new classes. It is also shown that the AIC module preserves the classification performance of old samples and resists catastrophic forgetting.

VI. Conclusions

This study proposes a novel FSCIL-EACA to adapt to the incremental recognition of few-shot images. Our model addresses the main challenges of FSCIL from the perspectives of embedding representation learning and classifier learning. First, we learn powerful embedding representations by introducing self-supervision and modulated attention training model on the base class. In the second stage, in order to achieve cross base and novel classes classification weights calibration and query data embeddings adaptation. We propose an incremental adaptive incremental classifier (AIC) program, including a hybrid relation projection module (HRP) and pseudo-incremental episode selection module (PES). Specifically, we project the old and novel class prototypes into the shared embedding space based on the AIC module and contextualize the prototypes to get the updated prototypes. Then, in order to adapt the query sample to the current task, we build a cross-mapping between the test data and the prototypes to update the test embeddings. Meanwhile, we use the PES module to simulate incremental episodes to enhance the scalability of the classifier. We conduct comparative and ablation experiments on four popular datasets, showing that the proposed method is general, suitable for medical image and natural image classification tasks, and has achieved superior recognition results.

Acknowledgement

This work was supported by the National Science Foundation of China (Grant No. 62176026), the Beijing

Natural Science Foundation (Grant No. M22009), and the Engineering Research Center of Information Networks, Ministry of Education.

References

[1] D. Pal, V. Bundele, R. Sharma, *et al.*, "Few-shot open-set recognition of hyperspectral images with outlier calibration network," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 2091–2100, 2022.

[2] S. Fang, X. B. Pan, S. M. Xiang, *et al.*, "Meta-MSNet: Meta-learning based multi-source data fusion for traffic flow prediction," *IEEE Signal Processing Letters*, vol. 28, pp. 6–10, 2021.

[3] Q. Wu, S. T. Miao, Z. L. Chai, *et al.*, "Fine-grained image classification with global information and adaptive compensation loss," *IEEE Signal Processing Letters*, vol. 29, pp. 36–40, 2022.

[4] D. W. Zhou, H. J. Ye, and D. C. Zhan, "Learning placeholders for open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 4399–4408, 2021.

[5] D. W. Zhou, Y. Yang, and D. C. Zhan, "Learning to classify with incremental new class," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2429–2443, 2022.

[6] S. A. Rebuff, A. Kolesnikov, G. Sperl, *et al.*, "iCaRL: Incremental classifier and representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 5533–5542, 2017.

[7] F. M. Castro, M. J. Marín-Jiménez, N. Guil, *et al.*, "End-to-end incremental learning," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 241–257, 2018.

[8] S. H. Hou, X. Y. Pan, C. C. Loy, *et al.*, "Lifelong learning via progressive distillation and retrospection," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 452–467, 2018.

[9] D. Isele and A. Cosgun, "Selective experience replay for lifelong learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, pp. 3302–3309, 2018.

[10] Y. Xiang, Y. Fu, P. Ji, *et al.*, "Incremental learning using conditional adversarial networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, pp. 6618–6627, 2019.

[11] S. H. Hou, X. Y. Pan, C. C. Loy, *et al.*, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 831–839, 2019.

[12] F. Zhu, X. Y. Zhang, C. Wang, *et al.*, "Prototype augmentation and self-supervision for incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 5867–5876, 2021.

[13] Z. C. Pan, X. H. Yu, M. H. Zhang, *et al.*, "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*, Waikoloa, HI, USA, pp. 6264–6273, 2023.

[14] F. Zhu, Z. Cheng, X. Y. Zhang, *et al.*, "Class-incremental learning via dual augmentation," in *Proceedings of the 35th Conference on Neural Information Processing Systems*, Online, pp. 14306–14318, 2021.

[15] X. Y. Tao, X. P. Hong, X. Y. Chang, *et al.*, "Few-shot class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 12180–12189, 2020.

[16] A. Kukleva, H. Kuehne, and B. Schiele, "Generalized and incremental few-shot learning by explicit learning and calibration without forgetting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp. 9000–9009, 2021.

[17] P. Mazumder, P. Singh, and P. Rai, "Few-shot lifelong learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, virtually, pp. 2337–2345, 2021.

[18] K. L. Chen and C. G. Lee, "Incremental few-shot learning via vector quantization in deep embedded space," in *Proceedings of the 9th International Conference on Learning Representations*, Virtual Event, Austria, 2021.

[19] C. Zhang, N. Song, G. S. Lin, *et al.*, "Few-shot incremental learning with continually evolved classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 12450–12459, 2021.

[20] G. Y. Shi, J. X. Chen, W. L. Zhang, *et al.*, "Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima," in *Proceedings of the 35th Conference on Neural Information Processing Systems*, Virtual, pp. 6747–6761, 2021.

[21] H. Shin, J. K. Lee, J. Kim, *et al.*, "Continual learning with deep generative replay," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 2994–3003, 2017.

[22] M. Y. Zhai, L. Chen, F. Tung, *et al.*, "Lifelong GAN: Continual learning for conditional image generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea, pp. 2759–2768, 2019.

[23] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 13, pp. 3521–3526, 2017.

[24] R. Aljundi, F. Babiloni, M. Elhoseiny, *et al.*, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 144–161, 2018.

[25] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, pp. 3987–3995, 2017.

[26] F. Sung, Y. X. Yang, L. Zhang, *et al.*, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1199–1208, 2018.

[27] W. B. Li, L. Wang, J. L. Xu, *et al.*, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 7253–7260, 2019.

[28] W. B. Li, J. L. Xu, J. Huo, *et al.*, "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, Honolulu, HI, USA, article no. 1060, 2019.

[29] H. J. Ye, H. X. Hu, D. C. Zhan, *et al.*, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8805–8814, 2020.

[30] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 4080–4090, 2017.

[31] O. Vinyals, C. Blundell, T. Lillicrap, *et al.*, "Matching networks for one shot learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain, pp. 3637–3645, 2016.

[32] C. Zhang, Y. J. Cai, G. S. Lin, *et al.*, “DeepEMD: Few-shot image classification with differentiable earth mover’s distance and structured classifiers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 12200–12210, 2020.

[33] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, pp. 1126–1135, 2017.

[34] S. K. Liu, A. J. Davison, and E. Johns, “Self-supervised generalisation with meta auxiliary learning,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, article no.150, 2019.

[35] A. Chavan, R. Tiwari, U. Bamba, *et al.*, “Dynamic kernel selection for improved generalization and memory efficiency in meta-learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 9851–9860, 2022.

[36] Y. Y. Zhao, L. N. Wang, Y. D. Tian, *et al.*, “Few-shot neural architecture search,” in *Proceedings of the 38th International Conference on Machine Learning*, Virtual Event, pp. 12707–12718, 2021.

[37] Y. J. Du, X. T. Zhen, L. Shao, *et al.*, “Hierarchical variational memory for few-shot learning across domains,” in *Proceedings of the Tenth International Conference on Learning Representations*, Virtual Event, 2022.

[38] H. B. Zhao, Y. J. Fu, M. T. Kang, *et al.*, “MgSvF: Multi-grained slow vs. fast framework for few-shot class-incremental learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press, 2021.

[39] D. W. Zhou, F. Y. Wang, H. J. Ye, *et al.*, “Forward compatible few-shot class-incremental learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 9036–9046, 2022.

[40] H. Liu, L. Gu, Z. X. Chi, *et al.*, “Few-shot class-incremental learning via entropy-regularized data-free replay,” in *Proceedings of the 17th European Conference on Computer Vision*, Tel Aviv, Israel, pp. 146–162, 2022.

[41] Z. X. Chi, L. Gu, H. Liu, *et al.*, “MetaFSCIL: A meta-learning approach for few-shot class incremental learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 14146–14155, 2022.

[42] X. L. Xu, S. S. Niu, Z. Wang, *et al.*, “Multi-feature space similarity supplement for few-shot class incremental learning,” *Knowledge-Based Systems*, vol. 265, article no. 110394, 2023.

[43] S. Gidaris and N. Komodakis, “Dynamic few-shot visual learning without forgetting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 4367–4375, 2018.

[44] K. H. Tang, J. Q. Huang, and H. W. Zhang, “Long-tailed classification by keeping the good and removing the bad momentum causal effect,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, article no.128, 2020.

[45] L. L. Jing and Y. L. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2021.

[46] Y. Z. Yang and Z. Xu, “Rethinking the value of labels for improving class-imbalanced learning,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, article no.1618, 2020.

[47] Z. W. Liu, Z. Q. Miao, X. H. Zhan, *et al.*, “Large-scale long-tailed recognition in an open world,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 2532–2541, 2019.

[48] X. L. Wang, R. Girshick, A. Gupta, *et al.*, “Non-local neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7794–7803, 2018.

[49] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6000–6010, 2017.

[50] D. J. Chen, H. Y. Hsieh, and T. L. Liu, “Adaptive image transformer for one-shot object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 12242–12251, 2021.

[51] H. Y. Zhang, M. Cissé, Y. N. Dauphin, *et al.*, “Mixup: Beyond empirical risk minimization,” in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada, 2018.

[52] H. Borgli, V. Thambawita, P. H. Smedsrud, *et al.*, “*HyperKvasir*, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy,” *Scientific Data*, vol. 7, no. 1, article no. 283, 2020.

[53] N. C. F. Codella, D. Gutman, M. E. Celebi, *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),” in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, USA, pp. 168–172, 2018.

[54] H. F. Wang, Z. F. Wang, M. N. Du, *et al.*, “Score-CAM: Score-weighted visual explanations for convolutional neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, pp. 111–119, 2020.

Ruru ZHANG was born in 1992. She is studying at Beijing University of Posts and Telecommunications for a doctorate in computer science and technology. Her interests include artificial intelligence and medical image processing. (Email: zrr@bupt.edu.cn)

Haihong E was born in 1982. She graduated from Beijing University of Posts and Telecommunications with a doctor’s degree in computer science and technology. At present, she is a Professor at Beijing University of Posts and Telecommunications. She has been engaged in research and teaching in the fields of big data, artificial intelligence, and cloud-native services for a long time.
(Email: ehaihong@bupt.edu.cn)

Meina SONG was born in 1974. She graduated from Beijing University of Posts and Telecommunications with a doctor’s degree in computer science and technology. She is a Professor at Beijing University of Posts and Telecommunications and director of the Information Network Engineering Research Center of the Ministry of Education. His main research direction is artificial intelligence and its application in the fields of finance and medicine.
(Email: mnsong@bupt.edu.cn)