Pattern Recognition 145 (2024) 109904

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Few-shot classification guided by generalization error bound ☆

Fan Liu [a], Sai Yang [b,*], Delong Chen [a], Huaxi Huang [c], Jun Zhou [d]

[a] College of Computer and Information, Hohai University, Nanjing 210098, China
[b] School of Electrical Engineering, Nantong University, Nantong, 226019, China
[c] Data61, CSIRO, Marsfield, 2122 NSW, Australia
[d] School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia

ARTICLE INFO

Keywords:
Few-shot classification
Generalization error bound
Self-supervised learning
Knowledge distillation

ABSTRACT

Recently, transfer learning has generated promising performance in few-shot classification by pre-training a backbone network on base classes and then applying it to novel classes. Nevertheless, there lacks a theoretical analysis on how to reduce the generalization error during the learning process. To fill this gap, we prove that the classification error bound on novel classes is mainly determined by the base-class generalization error, given the base-novel domain divergence and the novel-class generalization error produced by an incremental learner using novel samples. The novel-class generalization error is further decided by the base-class empirical error and the VC-dimension of the hypothesis space. Based on this theoretical analysis, we propose a Born-Again Networks under Self-supervised Label Augmentation (BANs-SLA) method to improve the generalization capability of classifiers. In this method, cross-entropy and supervised contrastive losses are simultaneously used to minimize the base-class empirical error in the expanded space with SLA. Afterward, BANs are adopted to transfer the knowledge sequentially across generations, which acts as an effective regularizer to trade-off the VC-dimension. Extensive experimental results have verified the effectiveness of our method, which establishes the new state-of-the-art performance on popular few-shot classification benchmark datasets.

1. Introduction

Biologically speaking, humans are innate to easily complete the recognition of massive natural and daily objects after observing just a few of samples. In contrast, even the latest advanced machine learning models like deep convolutional neural networks (CNNs) still heavily rely on a great quantity of high-quality labeled data to approach good performance. Thus, they are still struggling in many realistic scenarios, as annotating a large-scale data is usually very laborious and expensive. To bridge the significant gap between the CNNs and human intelligence, few-shot learning (FSL) [1,2] aiming at generalizing new concepts with a little supervision has rekindled an interest in many computer vision applications over years, such as image classification [3], object detection [4], semantic segmentation [5], and so on. Our work mainly focus on few-shot classification (FSC) [6–10], which attempts to train a model based on bass classes with sufficient annotated samples to predict unlabeled samples (query set), assuming only a few labeled samples (support set) are given per novel class.

For addressing this challenge issue, the most intuitive method is to make the machines to have the ability of learning to learn, as the way meta-learning paradigm has devoted to accumulating meta-knowledge for fast adaptation to novel classes with few labeled samples. Roughly speaking, meta-learning-based methods can be broadly divided into two types of optimization-based [6,11,12] and metric-based approaches [7,13–15]. Both sets of approaches adopt episodic training fashion to simulate real test environments. Despite great success with meta-learning, very recent studies [16–19] suspected that its sophisticated episodic training strategy is not the key factor for obtaining beneficial performance. Alternatively, they pre-trained a backbone network on the whole base dataset and then trained a traditional classifier for novel classes on the top of it. It is surprisingly that above transfer-learning paradigm has achieved competitive results compared with meta-learning. Under the transfer learning umbrella, the key to settling FSC tasks relies on a good embedding learned from base classes for the novel classes. Efforts to achieve this goal have been made

☆ This work was partially supported by National Nature Science Foundation of China (62372155), Joint Fund of Ministry of Education for Equipment Pre-research (8091B022123), Research Fund from Science and Technology on Underwater Vehicle Technology Laboratory (2021JCJQ-SYSJJ-LB06905), Key Laboratory of Information System Requirements, No: LHZZ 2021-M04, Water Science and Technology Project of Jiangsu Province under grant No. 2021063, Qinglan Project of Jiangsu Province.

* Corresponding author.
E-mail addresses: fanliu@hhu.edu.cn (F. Liu), yangsai@ntu.edu.cn (S. Yang), chendelong@hhu.edu.cn (D. Chen), Huaxi.Huang@csiro.au (H. Huang), jun.zhou@griffith.edu.au (J. Zhou).

https://doi.org/10.1016/j.patcog.2023.109904
Received 30 May 2023; Received in revised form 15 July 2023; Accepted 21 August 2023
Available online 26 August 2023

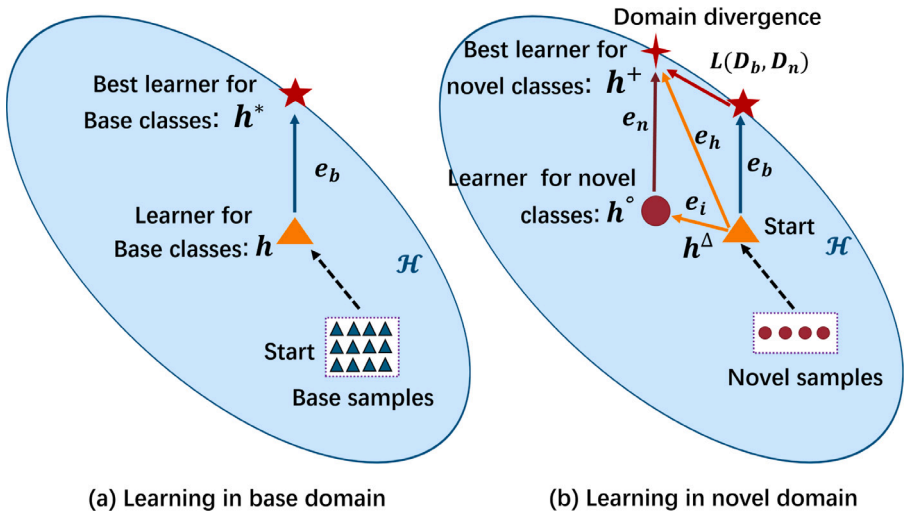


Fig. 1. (a) A learner $h \in \mathcal{H}$ is obtained with abundant base samples, resulting in the base-class generalization error of e_b. (b) The final learner $h^o \in \mathcal{H}$ is derived from a few novel samples based on h with difference h^Δ, yielding the generalization error on novel classes of e_n. e_h and e_i is the novel-classes generalization error respectively with h and h^Δ. e_n is composed of e_h and e_i. And e_h is bounded by e_b and $\mathcal{L}(D_b, D_n)$. So, the generalization error bound e_{bound} for the novel classes is formulated as: $e_{bound} = e_b + \mathcal{L}(D_b, D_n) + e_i$.

from different perspectives, for example, learning a general-purpose feature extractor with the manifold technique [18], improving model transferability via expanded margin in loss function [19], and obtaining invariant features by self-supervised learning [9]. Moreover, the self-distillation technique [17] is also proven to be able to improve the performance of FSC. The diversity of the above methods naturally raises an interesting question: *How to enforce a model pre-trained on the base classes to act well on novel classes?* Or alternatively, *is there a general guideline for FSC to obtain good performance on unseen novel classes?*

We attempt to answer these questions by conducting a theoretical study on the generalization error bound of FSC, which unveils the impact of errors produced in each stage of transfer learning. Fig. 1 summarizes the core idea of our theorem. As shown in Fig. 1(a), a mapping function $h \in \mathcal{H}$ is learned from the base set with abundant training samples. In Fig. 1 1(b), the final mapping function $h^o \in \mathcal{H}$ is obtained with a few samples of novel classes based on h with difference h^Δ. In this case, the generalization error on the novel classes is composed of the generalization error from both h and h^Δ. According to the domain adaptation learning theory [20], the former is bounded by the generalization error on the base set and the domain divergence between the base-class and the novel-class. Therefore, the classification error bound on the novel classes e_{bound} is related to: **(1)** e_{b}: the base-class generalization error with h, **(2)** $\mathcal{L}(D_{\mathrm{b}}, D_{\mathrm{n}})$: the base-novel domain divergence, and **(3)** e_{i}: the novel-class generalization error with h^Δ. The above three terms can be concluded to be the following formula of $e_{bound} = e_b + \mathcal{L}(D_b, D_n) + e_i$. Decreasing the above three terms leads to a better FSC classifier for the novel class. Wherein, the pre-training stage is responsible for the first term, which is determined by the empirical error and the VC-dimension of the hypothesis space according to the statistical learning theory [21]. However, current transfer learning based FSC methods [17–19] mainly focused on leveraging various regularization techniques to avoid over-fitting the base set, ignoring the empirical error on the base set which is also a vital term for the final error bound. Moreover, these methods usually suffer from tuning many parameters to balance loss terms in the final loss function.

To solve the above problems, we instantiate a new algorithm named Born Again Networks under Self-supervised Label Augmentation (BANs-SLA) according to the proposed theorem. BANs-SLA is designed to directly minimize the empirical error on the base set by only tuning the weight of the Kullback–Leibler (KL) loss term. In specific, SLA [22] is exploited to expand the original label space. This can avoid the conflict between the self-supervised and the originally supervised tasks [23], whilst controlling the VC-dimension of the learning model implicitly by manipulating the data complexity with label augmentation. With SLA, the Cross-Entropy (CE) and Supervised Contrastive (SC) losses [24] are simultaneously used to minimize the empirical error. We then adopt an effective regularization strategy of BANs [25] to transfer the knowledge sequentially across generations with the aim of further decreasing the generalization error. In summary, the contributions of this paper are:

- To our knowledge, this is the first theoretical study on FSC in the context of the transfer learning paradigm. To this end, we propose a generalization error bound theorem to guide FSC.
- Following our theorem, we propose a Born-Again Networks under Self-supervised Label Augmentation (BANs-SLA) method for FSC, in which the joint learning with the CE and SC losses targets minimizing the empirical error, while the strategies of SLA and BANs trade-off the VC-dimension of the hypothesis space.
- We conduct extensive experiments on multiple benchmarks to demonstrate the effectiveness of the BANs-SLA method, which has experimentally validated the proposed theorem.

2. Related work

2.1. Few-shot classification

Generally, the goal of FSC is to endow AI systems with the ability of generalizing new concepts under low-data regime. To this end, meta-learning seeks to accumulate meta-knowledge for fast adaptation by organizing the training into a series of episodes. According to knowledge type, it is totally grouped into two categories of optimization-based and metric-based methods. The former performs an explicit bi-level optimization process to possess a meta-learner for fast optimization within a few steps. The seminal works including MAML [6] and its variant [26] tried to learn initialization parameters to provide a good start point for unseen tasks. Except for the initialization parameters, Meta-Learner LSTM [27] focused on learning the updating rules of an optimizer. For further enhancing the generalization, recent works of this type like MetaOptNet [28] and MeTAL [29] respectively explored hinge loss and task-adaptive loss to substitute the common CE loss in the inner-loop optimization. Compared with the first type, metric-learning based approaches have shown to be more prominent in settling FSC with attractive simplicity and effectiveness. This type of approaches follow the idea of jointly learning a good feature and metric to differentiate samples per class. The earlier well-known models including Matching Networks [30], Prototypical Networks [7] and Relational Networks [8]

have been successively proposed to form the generic FSC framework. Afterwards, many literature improved upon them mainly from two respects, i.e. attaining a powerful feature extractor and designing a richer similarity metric. For example, RENet [31], MFS [32] and TPMN [33] resorted to attention mechanisms to capture distinctive features, while DeepEMD [34], DAN [35] and HGNN [36] respectively leveraged Wasserstein distance, dynamic filter and graphical model to calculate similarity based on local feature descriptors.

Recently, a handful of works cast questioning the efficacy of the episodic training strategy, and forewent such practice with performing training on the whole base dataset instead. The early influence work appeared in [16], where Baseline and Baseline++ pre-trained CNNs with a linear or cosine classifier from scratch using CE loss and then fine-tuned the classifier weights for novel classes. Evidently, this kind of methods follow transfer-learning paradigm, which largely hinge on training a good feature extractor. For this, Neg-Cosine [19] and S2M2 [18] respectively employed negative-margin softmax loss and manifold mixup to learn generalization feature representations. Besides, as our method is most related to many works using self-supervised learning and knowledge distillation to enhance feature representations, we will review them comprehensively in Sections 2.2 and 2.3. Moreover, as the classifier used in pre-training is thrown away, current methods usually reuse a traditional classifier to serve for the novel classes. For example, DC-LR [37] generated more features according to a calibrated Gaussian distribution to train a logistic regression (LR) classifier to predict query samples. CCF [38] generated category-correlated features to train linear classifier for novel classes. DeepBDC [39] inserted a BDC module to enhance feature representation of backbone network and also leveraged LR to classify novel classes. The recent SGI [40] designed an improved convolution structure named Self-Guided Information Convolution to extract discriminative features. Although various methods have constantly refreshed the FSC performance, it still lacks of a rule to direct pre-training.

2.2. Self-supervised learning

Recent years have witnessed a rapid development of self-supervised learning that aims to learn generalized feature representations without any label by elaborately designing a set of proxy tasks [41]. More recently, several studies [42,43] investigated the underlying mechanisms of self-supervised learning from various angles. Remarkably, [43] combined self-supervised learning and supervised learning to get better classification performance. In line with the above studies, FSC can be done by a linear combination of self-supervised learning and supervised learning models under either meta-learning or transfer learning setting. For example, IEPT [44] and CC+rot [45] adopted the meta-learning to introduce a rotation prediction task to enhance the transferability of the backbone network, while CSIV [9] introduced rotation prediction and instance discrimination as auxiliary tasks to improve the generalization of features under the umbrella of transfer learning. However, [23] suggested that the direct combination of two learning paradigms may lead to conflicts between tasks. To circumvent this problem, [22] proposed a Self-supervised Label Augmentation (SLA) method to learn a single unified task with respect to the joint distribution of the original and self-supervised labels. On this foundation, we propose to simultaneously use cross-entropy and supervised contrastive losses to implement this unified classification task more effectively.

2.3. Knowledge distillation

Knowledge distillation has recently been adopted for model compression by learning a small student network from a large teacher network with knowledge transfer. In the absence of a high-quality teacher network, self-distillation [46] and co-distillation [47] were proposed for knowledge transfer in FSC. RFS [17] and SKD [48] used the born-again strategy to sequentially transfer the knowledge through multiple generations, differing in the way on how to train a good primary model. PAL [49] optimized a teacher network by using the supervised contrastive loss and forcing the student network to align their logit and feature to the teacher network. BML [50] leveraged mutual learning between the transfer learning based and meta-learning based methods. Unlike these methods, our method employs self-distillation for FSC under SLA.

3. The proposed method

3.1. Theory foundation

Under the transfer learning based FSC setting, an available dataset D is divided into three disjoint sets of the base set, validation set and novel set. We respectively denote the base set and the validation set as D_b and D_{val}. The novel set is organized into a series of episodes, and we denote the subset as D_n in each episode, which is composed of N classes with several K labeled samples. As discussed in the Introduction Section, transfer learning based methods aim to pre-train a backbone network on D_b with validation on D_{val} for dealing with FSC tasks on D_n. An interesting question here is: *how to fully exploit D_b to obtain good performance on D_n?* We attempt to answer this question by investigating the underlying mechanism behind the transfer learning from the view of the generalization error bound.

For simplicity, we first make some basic definitions in the case of binary classification (i.e., $N = 2$). Please note that our theory can be naturally extended to multi-class classification (i.e., $N > 2$). Following [20], two special domains of base classes and novel classes are considered in our work, denoted as $\{D_b, f_b\}$ and $\{D_n, f_n\}$, wherein f_b and f_n respectively represents the label function of D_b and D_n. The size of the labeled samples on D_b and D_n is L and K, respectively. Let $\mathcal{H}$ be a hypothesis space, and a particular mapping function $h \in \mathcal{H}$ is learned using the base set. In the novel domain, the mapping function $h^o \in \mathcal{H}$ is learned from h with difference $h^\triangle$ using the novel samples. Then the expected error with h^o on D_n is approximated as:

$$e_n(h^o) \approx e_n(h) + e_n(h^\triangle). \tag{1}$$

Then the expected error with h and $h^\triangle$ respectively on D_b and D_n are:

$$\begin{aligned} e_b(h) &= e_{D_b}(h, f_b) = E_{x\in D_b}[\left|h(x) - f_b(x)\right|], \\ e_n(h^\triangle) &= e_{D_n}(h^\triangle, f_n) = E_{x\in D_n}[\left|h^\triangle(x) - f_n(x)\right|]. \end{aligned} \tag{2}$$

We assume that there exists a significant domain shift between the base and novel classes in the FSC task. Thus, we use the variation [20] to measure the divergence between the distributions between D_b and D_n:

$$\mathcal{L}(D_b, D_n) = 2 \sup_{B\in\mathcal{B}} \left| Pr_{D_b}(B) - Pr_{D_n}(B) \right|, \tag{3}$$

where $\mathcal{B}$ is a subset of the union of D_b and D_n. $Pr(B)$ represents the probability of set B.

Theorem 1 (*FSC Generalization Error Bound*). *Let $\mathcal{H}$ be a hypothesis space, and υ be the Vapnik–Chervonenkis (VC) dimension of $\mathcal{H}$. For every $h^o, h, h^\triangle \in \mathcal{H}$ and $\eta \in [0, 1]$, the expected error bound on D_n holds the following relationship with probability at least 1-η:*

$$e_n(h^o) \le \underbrace{\hat{e}_b(h) + \sqrt{\frac{\upsilon(ln\frac{2L}{\upsilon}+1) - ln\frac{\eta}{4}}{L}}}_{\text{generalization error on } D_b \text{ with } h} + \underbrace{\mathcal{L}(D_b, D_n)}_{D_b - D_n \text{ divergence}} + \underbrace{\hat{e}_n(h^\triangle) + \sqrt{\frac{\upsilon(ln\frac{2K}{\upsilon}+1) - ln\frac{\eta}{4}}{K}}}_{\text{generalization error on } D_n \text{ with } h^\triangle} + \lambda, L \gg K,$$

where L is the size of labeled samples on D_b, K is the size of labeled samples on D_n, $\hat{e}_b(h)$ is the empirical error on D_b, $\hat{e}_n(h^\triangle)$ is the empirical error on D_n, and λ is a constant.

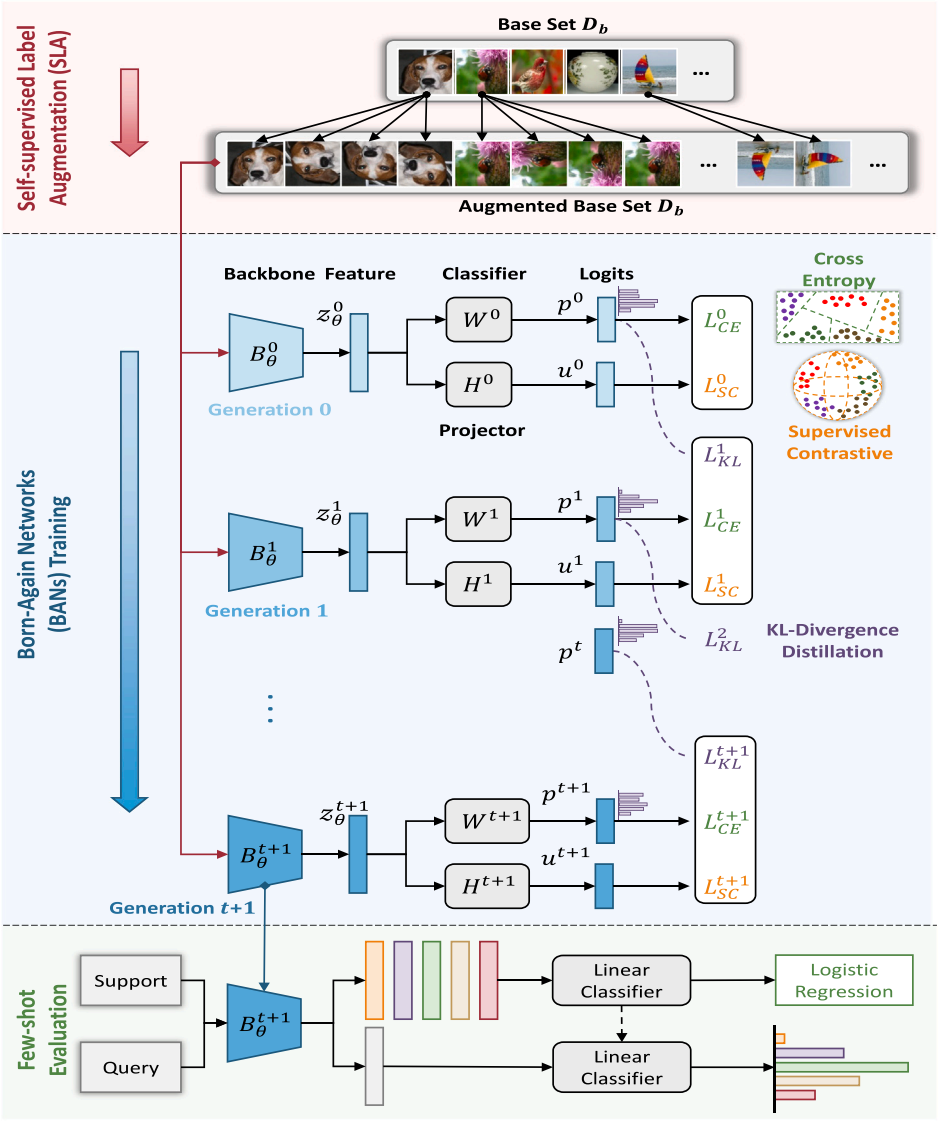


Fig. 2. An overview of the proposed framework. It firstly expands the label space of base set D_b with Self-supervised Label Augmentation (SLA) and then undertakes sequential knowledge distillation across different generations (i.e., Generation 0, Generation 1, ..., Generation t+1) with Born-Again Networks (BANs). The student network in the last generation is deployed to perform few-shot evaluations.

The proof is provided in the Supplementary Material.

Remark 1. Theorem 1 tells the error bound on D_n is determined by three terms: **(1)** the generalization error on D_b with h, **(2)** the generalization error on D_n with $h^{\triangle}$, and **(3)** the domain divergence between D_b and D_n. Therefore, the basic rule to guide FSC is to minimize the above three terms, so that the main objective of FSC (i.e., decreasing $e_n(h^o)$) can be accomplished. Under the transfer learning based FSC setting, the pre-training process is mainly responsible for the first term (minimizing $e_b(h)$), given the other two terms. This can well explain why various regularization techniques discussed in the Introduction Section are effective, i.e., they lower the generalization error on D_b via obtaining robust features.

3.2. Born-again networks under self-supervised label augmentation

3.2.1. Overview

Our method mainly focus on $e_b(h)$ in Theorem 1, which is determined by both $\hat{e}_b(h)$ and υ of the hypothesis space. Accordingly, a powerful FSC model can be obtained by selecting the most suitable techniques to simultaneously focus on $\hat{e}_b(h)$ and υ. For the first term $\hat{e}_b(h)$, existing methods mainly learn a classification hyperplane by the Cross-Entropy (CE) loss function. Here we further construct a hypersphere by employing the Supervised Contrastive (SC) loss function to refine the classification boundary. For the second term υ, we propose an algorithm named Born-Again Networks under Self-supervised Label Augmentation (BANs-SLA), in which iterative Knowledge Distillation (KD) and transformation-based training set expansion serve as strong regularizations to trade-off υ of the learning model. An overview of our method is shown in Fig. 2, and a flow description of BANs-SLA is given in Algorithm 1. In our learning model, the image processing module denoted as $T(\cdot)$ performs SLA, which transforms the images and re-annotates them with the newly augmented class labels. The backbone network is denoted as $B_\theta(\cdot)$ parameterized by θ, which maps images into a d-dimensional feature space. The linear classier is denoted as $C_w(\cdot)$ with parameter matrix $W \in \mathcal{R}^{d\times MC}$, which classifies the images and their transformations into one of the augmented categories. The projector is denoted as $P_h(\cdot)$ with parameter matrix $H \in \mathcal{R}^{d\times Q}$ being

used to project the image feature into a Q-dimensional hypersphere space.

3.2.2. Primitive training with self-supervised label augmentation

During the primitive training in Generation 0, the learning model is denoted to be $I_0 = \{B_\theta^0, C_w^0, P_h^0\}$. Given a batch of L images randomly sampled from D_b, let x_i be any image, $y_i \in \{1, 2, \ldots, C\}$ be its original label, where C is the total number of base classes. $T(\cdot)$ applies M kinds of transformation to each image, resulting in ML image samples. In the meantime, the label space has also been expanded by M times, i.e., the label of x_i turns to be $\hat{y}_i \in \{1, 2, \ldots, MC\}$. Feed x_i into B_θ^0 to produce a d-dimensional feature, which is formulated as $z_i^0 = B_\theta^0(x_i) \in \mathscr{R}^d$. The features then go through the linear classier and its corresponding softmax layer to output the predicted probability, of which the jth component can be written as:

$$P_{ij}^0 = \sigma\left(C_w^0\left(z_i^0\right)\right) = \frac{exp({z_i^0}^T w_j)}{\sum_{j=1}^{MC} exp({z_i^0}^T w_j)}, \tag{4}$$

where σ is the softmax function, $w_j \in \mathscr{R}^d$ is the jth classification weight vector of parameter matrix $W^0 \in \mathscr{R}^{d\times MC}$. Then the cross-entropy classification loss under SLA is:

$$L_{CE}^0\left(\theta^0, W^0\right) = -\sum_{i=1}^{ML}\sum_{j=1}^{MC} \hat{y}_{ij} log p_{ij}^0, \tag{5}$$

where $\hat{y}_{ij}$ is the jth component of label $\hat{y}_i$.

Additionally, the image feature z_i^0 is also fed into the projector to output a Q-dimensional normalized feature $u_i^0 \in \mathscr{R}^Q$, which is written as:

$$u_i^0 = \left\| P_h^0(z_i^0) \right\| = \left\| {z_i^0}^T H^0 \right\|, \tag{6}$$

where $H^0 \in \mathscr{R}^{d\times Q}$ is the parameter matrix. Assume the whole set of samples in the given batch consist of a set of $A\left(u_i^0\right)$. Let u_i^0 be an anchor point to index all the positive samples that have the same augmented label with it. The positive samples construct the set of $P\left(u_i^0\right)$. Then the supervised contrastive loss is adopted to pull the samples of the same class together while pushing apart samples from different classes in the augmented label space, as follows:

$$L_{SC}^0(\theta^0, H^0) = \sum_{i\in A(u_i^0)} -log\{\frac{1}{|P(u_i^0)|} \sum_{p\in P(u_i^0)} \frac{exp(u_i^0 \cdot u_p^0/\tau)}{\sum_{a\in A(u_i^0)} exp(u_a^0 \cdot u_p^0/\tau)}\}, \tag{7}$$

where τ is a scalar temperature parameter, $|P(u_i^0)|$ is the cardinality of $P\left(u_i^0\right)$, u_p^0 is the pth sample from $P(u_i^0)$, u_a^0 is the ath sample from $A\left(u_i^0\right)$. Then the overall loss function for primitive training with SLA is :

$$L_{Gen0}\left(\theta^0, W^0, H^0\right) = L_{CE}^0\left(\theta^0, W^0\right) + \alpha L_{SC}^0\left(\theta^0, H^0\right). \tag{8}$$

As CE and SC have the identical function of minimizing the empirical error, α is empirically set to be 1. Gradient descent is used to update the parameters to approach the teacher network for the first generation.

3.2.3. Born-again networks

The next task is to depict the learning process from the tth ($t > 0$) generation to $(t+1)$th generation. To this end, the student network approximates the joint distribution of the original and self-supervised labels using CE and SC losses, whilst learning information from the teacher network of the former generation. For convenience, we denote the learning model of the teacher network and the student network as $I_t = \{B_\theta^t, C_w^t, P_h^t\}$ and $I_{t+1} = \{B_\theta^{t+1}, C_w^{t+1}, P_h^{t+1}\}$.

Given a batch of L images randomly sampled from D_b, any image x_i after SLA is simultaneously fed into the teacher network and the student network. The output features from each backbone network are defined as $z_i^t = B_\theta^t\left(x_i\right) \in \mathscr{R}^d$ and $z_i^{t+1} = B_\theta^{t+1}\left(x_i\right) \in \mathscr{R}^d$. Then the features pass through each classifier to output prediction probability $p^t\left(x_i\right) = \sigma\left(C_w^t\left(z_i^t\right)\right)$ and $p^{t+1}\left(x_i\right) = \sigma\left(C_w^{t+1}\left(z_i^{t+1}\right)\right)$. At present, the optimization function for the student network of the $(t+1)$th generation with CE loss is:

$$L_{CE}^{t+1}\left(\theta^{t+1}, W^{t+1}\right) = -\sum_{i=1}^{ML}\sum_{j=1}^{MC} \hat{y}_{ij} log p^{t+1}\left(x_i\right). \tag{9}$$

The feature from the projector of the student network is $u_i^{t+1} = p_h^{t+1}(z_i^{t+1})$. Similar to the primitive training in the former part, let u_i^{t+1} be an anchor point to define its whole set of samples $A\left(u_i^{t+1}\right)$ and let the positive sample set to be $P\left(u_i^{t+1}\right)$. Then the SC loss in the $(t+1)$th generation is :

$$L_{SC}^{t+1}(\theta^{t+1}, H^{t+1}) = \sum_{i\in A(u_i^{t+1})} -log\{\frac{1}{|P(u_i^{t+1})|} \sum_{p\in P(u_i^{t+1})} \frac{exp(u_i^{t+1} \cdot u_p^{t+1}/\tau)}{\sum_{a\in A(u_i^{t+1})} exp(u_a^{t+1} \cdot u_p^{t+1}/\tau)}\}, \tag{10}$$

where τ is a scalar temperature parameter, $|P(u_i^{t+1})|$ is the cardinality of $P\left(u_i^{t+1}\right)$, u_p^{t+1} is the pth sample from $P(u_i^{t+1})$, and u_a^{t+1} is the ath sample from $A\left(u_i^{t+1}\right)$.

During the process of knowledge distillation, the student network mainly assimilates softened output knowledge from the teacher network. Here, the Kullback–Leibler (KL) divergence of the prediction probability is chosen to measure the output information. The KL distance from the teacher network to the student network is computed as:

$$D_{KL}^{t+1}\left(\theta^{t+1}, W^{t+1}\right) = \sum_{i=1}^{ML} p^{t+1}(x_i) \frac{p^{t+1}(x_i)}{p^t(x_i)/\epsilon}, \tag{11}$$

where ϵ is a hyper-parameter of temperature. Then the final loss function for optimizing the student network in the $(t+1)$th generation is formulated as:

$$L_{Gen^{t+1}}\left(\theta^{t+1}, W^{t+1}, H^{t+1}\right) = L_{CE}^{t+1}\left(\theta^{t+1}, W^{t+1}\right) + \alpha L_{SC}^{t+1}(\theta^{t+1}, H^{t+1}) + \beta D_{KL}^{t+1}\left(\theta^{t+1}, W^{t+1}\right), \tag{12}$$

where β is the weight of the KL loss term to be tuned.

Algorithm 1: Born-Again Networks under Self-supervised Label Augmentation

Input: Base dataset $D_b = \{(x_i, y_i)\}_{i=1}^L$, augmentation module $T(\cdot)$, backbone network $B_\theta(\cdot)$, linear classifier $C_w(\cdot)$, projector $P_h(\cdot)$
Output: A well-trained backbone network
Gen0: Primitive Training with SLA
for *numbers of training epochs* **do**
  Sample a mini-batch with any image of (x_i, y_i);
  Feed x_i into $T(\cdot)$ and B_θ^0 to obtain featurez_i^0 ;
  Pass z_i^0 through C_w^0 to get the output probability;
  Pass z_i^0 through P_h^0 to get the projection feature;
  Calculate optimization loss via Eq. (8);
  Update parameters of θ^0, W^0, H^0 using SGD;
end
Gen(t+1): Bon-Again Networks (BANs)
for *numbers of training epochs* **do**
  Sample a mini-batch with any image of (x_i, y_i);
  Feed x_i into I^t and I^{t+1} to obtain features and probability output;
  Calculate KL loss between I^t and I^{t+1};
  Compute overall loss for I^{t+1} via Eq. (12);
  Update parameters θ^{t+1}, W^{t+1}, H^{t+1} using SGD for I^{t+1};
end

Table 1
Test accuracy (%) of each component of our method under 5-way 1-shot and 5-shot tasks on benchmark datasets.

Method	Backbone	miniImageNet		CIFAR-FS		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Gen0_CE	ResNet12	67.37 ± 0.42	84.36 ± 0.28	73.96 ± 0.46	88.18 ± 0.31	78.26 ± 0.42	92.00 ± 0.20
Gen0_SC	ResNet12	65.34 ± 0.43	81.16 ± 0.32	69.85 ± 0.47	82.40 ± 0.36	65.61 ± 0.49	86.40 ± 0.20
Gen0_CE+SC	ResNet12	69.05 ± 0.43	84.87 ± 0.29	77.45 ± 0.45	88.64 ± 0.33	80.30 ± 0.42	92.38 ± 0.20
Gen1	ResNet12	70.40 ± 0.44	85.31 ± 0.27	77.98 ± 0.45	89.78 ± 0.31	83.17 ± 0.40	93.75 ± 0.19

Table 2
Test accuracy (%) of our method with and without SLA under 5-way 1-shot and 5-shot tasks on benchmark datasets.

SLA	Backbone	miniImageNet		CIFAR-FS		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
×	ResNet12	63.34 ± 0.46	80.28 ± 0.32	74.16 ± 0.50	85.16 ± 0.35	79.17 ± 0.47	90.11 ± 0.31
✓	ResNet12	69.05 ± 0.43	84.87 ± 0.29	77.45 ± 0.45	88.64 ± 0.33	80.30 ± 0.42	92.38 ± 0.20

Table 3
Test accuracy (%) of our method with and without BANs under 5-way 1-shot and 5-shot tasks on benchmark datasets.

BANs	Backbone	miniImageNet		CIFAR-FS		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
×	ResNet12	69.05 ± 0.43	84.87 ± 0.29	77.45 ± 0.45	88.64 ± 0.33	80.30 ± 0.42	92.38 ± 0.20
✓	ResNet12	70.40 ± 0.44	85.31 ± 0.27	77.98 ± 0.45	89.78 ± 0.31	83.17 ± 0.40	93.75 ± 0.19

3.3. Few-shot evaluation

After pre-training, the student network in the last generation is deployed to implement few-shot evaluation, in which the projector and linear classifier are removed and the backbone network is frozen to play as a feature extractor. In each FSC task, B_θ^{t+1} is used to extract features for support and query samples in D_n. A plain classifier of logistic regression $g_\phi(\cdot)$ with parameter ϕ is trained with support features to predict the label for each query image.

4. Experiments

4.1. Datasets

miniImageNet has 100 classes with 600 images per class. These classes are split into 64, 16, and 20 respectively for the base, validation, and novel sets [30]. **tiredImageNet** contains 608 classes with an average number of 1281 images per class. The images are split into 351, 97, and 160 classes respectively for the base, validation, and novel sets [51]. **CIFAR-FS** contains 100 classes with 600 images per class. The total classes are divided into 64, 16, and 20 for the base, validation, and novel sets [28]. **Caltech-UCSD Bird-200-2011 (CUB)** contains 200 bird species with a total number of 11,788 images. The images are divided into 100, 50, and 50 classes respectively for the base, validation, and novel sets [52].

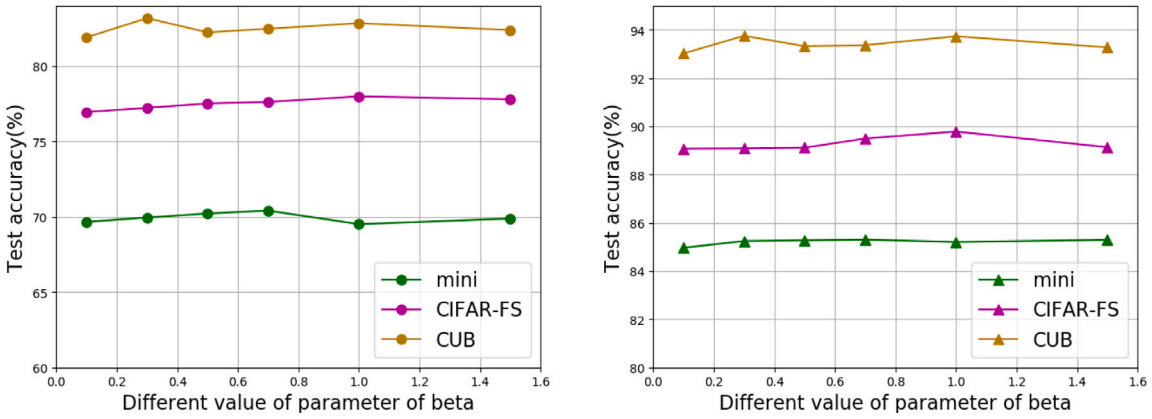
4.2. Implementation details

For a fair comparison with previous methods, ResNet12 is adopted as the backbone network in our method. It consists of 4 residual blocks with 640 filers in the last block, resulting in a 640-dimensional global feature for each input image. The projector network is a multi-layer perceptron with only a single linear layer of size 128. For all the experiments, the SGD with a momentum of 0.9 and a weight decay of 5e−4 is chosen as the optimizer. The training epoch number is 130 and the batch size in each epoch is 32. For miniImageNet, tiredImageNet and CIFAR-FS, the initial learning rate is set to 0.025 and decreased by 0.2 at the 70th and 100th epochs. For CUB, the initial learning rate is set to 0.1 and decayed by 0.2 for every 15 epochs after the 75th epoch. The temperature parameter in the SC loss and the KL loss is set to 0.1 and 4, respectively. During the evaluation phase, we perform 5-way 1-shot and 5-shot FSC tasks on all the datasets. In each case, we implement a meta-test with 2000 episodes, in which each episode randomly samples 15 query images from each novel class. The results are finally reported as mean classification accuracy over all the episodes and its corresponding 95% confidence intervals.

4.3. Ablation studies

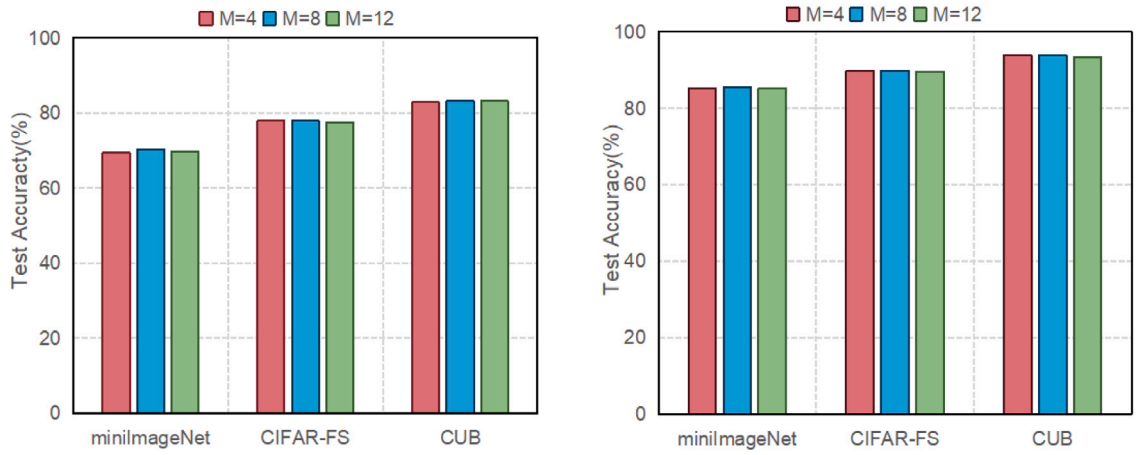
We conduct ablation studies to investigate the effect of individual components in two aspects. Firstly, the training of the teacher network in the original generation involves the joint learning of CE and SC losses. Thus, we analyze the effectiveness of their linear combination versus being used alone respectively. This experiment results in three methods denoted as Gen0_CE, Gen0_SC, and Gen0_CE+SC. Secondly, knowledge transfer from the original generation Gen0_CE+SC to the first generation Gen1 is another important component. The test accuracy of each component of our method under 5-way 1-shot and 5-shot tasks on popular benchmark datasets is shown in Table 1. From the results, we can see that: (1) On all the datasets, Gen0_CE and Gen0_SC exhibit different performances in both 1-shot and 5-shot tasks. For example, Gen0_CE exceeds Gen0_SC by a large margin, up to at least 6% in both 1-shot and 5-shot settings on the CUB dataset. This result illustrates that CE and SC losses have different properties in dealing with the supervised learning task, which implies that their joint learning may be beneficial. (2) Gen0_CE+SC outperforms both Gen0_CE and Gen0_SC in all the cases, which tells that the combination of CE and SC losses is more effective than them being used alone under SLA. This observation validates that joint learning with CE and SC losses can further minimize the empirical error. (3) The performance of Gen1 exceeds Gen0_CE+SC under 1-shot and 5-shot tasks on all the datasets, which demonstrates that the strategy of BANs is still very effective under SLA.

Otherwise, the proposed method mainly consists of two components, i.e. SLA and BANs. Therefore, we investigate the important of each component. On the one hand, we report the test accuracy of our method with and without SLA under 5-way 1-shot and 5-shot tasks in Table 2. As shown with the results, we find the SLA can significantly increase the performance of our method, illustrating the technique of SLA is a very effective technique to improve the generalization of the backbone network. On the other hand, we report the test accuracy of our method with and without BANs under 5-way 1-shot and 5-shot tasks in Table 3. As shown with the results, we can see that the performance with BANs outperforms the one without BANs, stating BANs can significant contribution in our method.



(a) 5-way 1-shot (b) 5-way 5-shot

Fig. 3. Test accuracy (%) of different parameter values under 5-way 1-shot and 5-shot on FSC datasets.



(a) 5-way 1-shot (b) 5-way 5-shot

Fig. 4. Test accuracy (%) of different transformations under 5-way 1-shot and 5-shot on miniImageNet, CIFAR-FS & CUB.

4.4. Hyper-parameter analysis

As shown in Eq. (12), the overall loss is mainly composed of three terms, i.e., CE, SC and KL loss. Among them, CE and SC share the same aim of minimizing the empirical error, which have the identical weight of 1. Consequently, there only one hyper-parameter of β left to be tuned. We vary the value of β between [0, 1.5] and show the accuracy curves under different values in Fig. 3. From the results, we can see that the highest performance is respectively reported at $\beta = 0.7$, 1 and 0.3 on miniImageNet, CIFAR-FS and CUB. It is worth noting that the discrepancy between the maximum and minimum is marginal, which indicates that it is easy to tune the only hyper-parameter in our method.

4.5. The number of the transformations

SLA which is realized by rotating images under different scales, is an important component in our method. We mainly investigate four rotation angles and three scales to get the transformations number of 4, 8 and 12, respectively. The test accuracy under different transformations on popular benchmark datasets is shown in Fig. 4. From the results, we can observe that on miniImageNet and CUB, the highest performance is obtained when M=8. Increasing more transformations does not bring any further performance improvement. on CIFAR-FS, there is no big performance difference among different transformations. Thus, we set the transformation number to be 8 in our all experiments.

4.6. Time complexity analysis

Our method implements knowledge distillation across different generations. The total number of generations is assumed to be T. In each generation, the training is related to three factors, i.e. the number of training epoch E, the batch size in each epoch L, the number of transformations in SLA M. Then our method has the time complexity $O(TELM)$. Given the number of M and L, the time complexity mainly depends on the number of generations. As shown in Fig. 5, we investigate the FSC performance under different generations. From the results, we can see there is a big performance leap from Gen0 to Gen1, and the performance curves tend to be stable after Gen1. So our method can achieve good performance when the generation number is only set to 1, not bringing much extra computation time.

4.7. Robust feature extraction and t-SNE visualization

In this subsection, we discuss the robustness of our method's feature extraction, which is crucial for handling the diversity of novel class features. To illustrate this, we perform a t-SNE visualization of test images, randomly sampling 5 classes and 200 images per class from the novel dataset of miniImagenet. The visualization results of Gen0_CE, Gen0_SC, Gen0_CE+SC, and Gen1 are shown in Fig. 6. The results reveal that the 5 classes are better separated in the feature representation

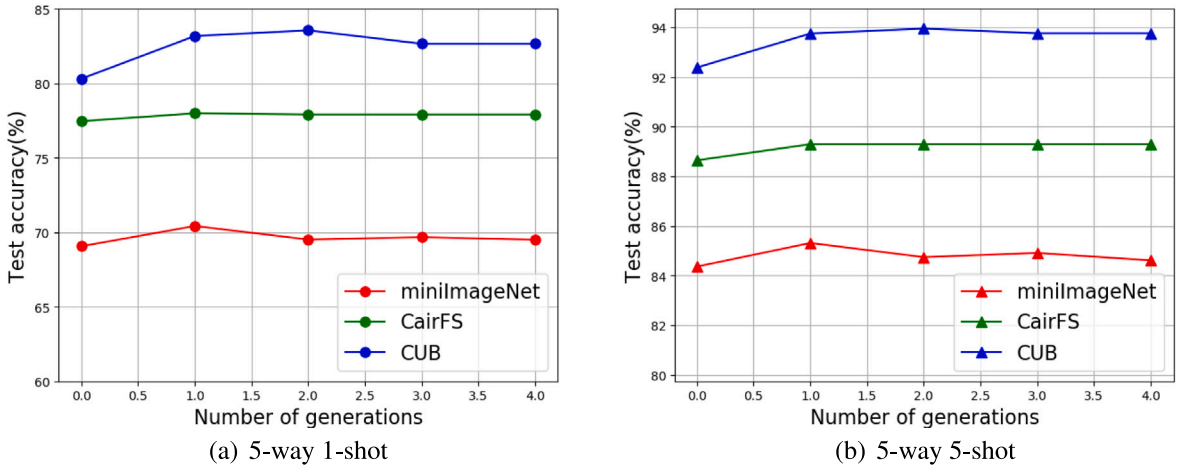


Fig. 5. Test accuracy (%) of different generations under 5-way 1-shot and 5-shot on miniImageNet, CIFAR-FS and CUB.

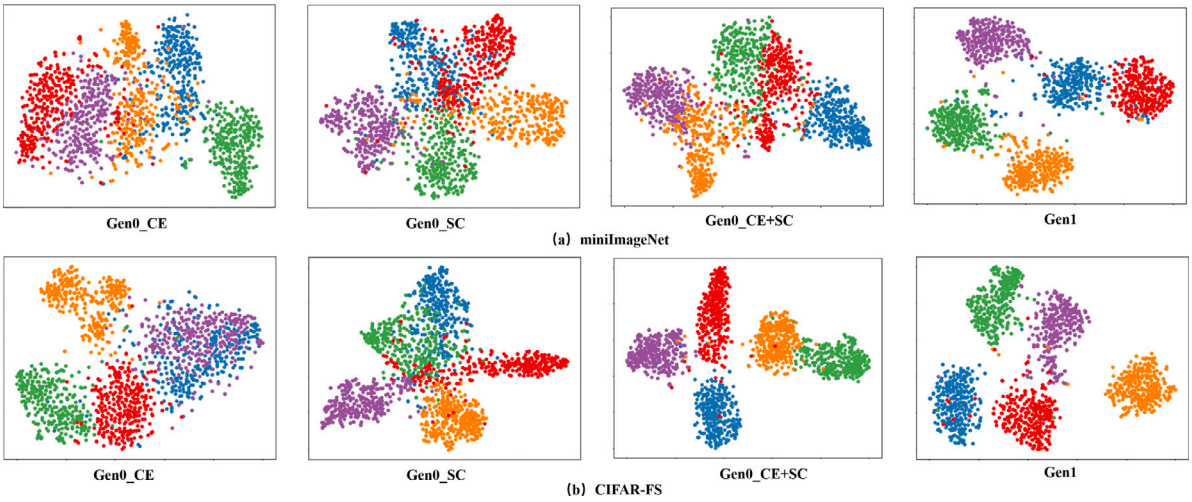


Fig. 6. t-SNE visualization of support features of novel classes extracted by the backbone network pre-trained with Gen0_CE, Gen0_SC, Gen0_CE+SC, Gen1 on miniImageNet.

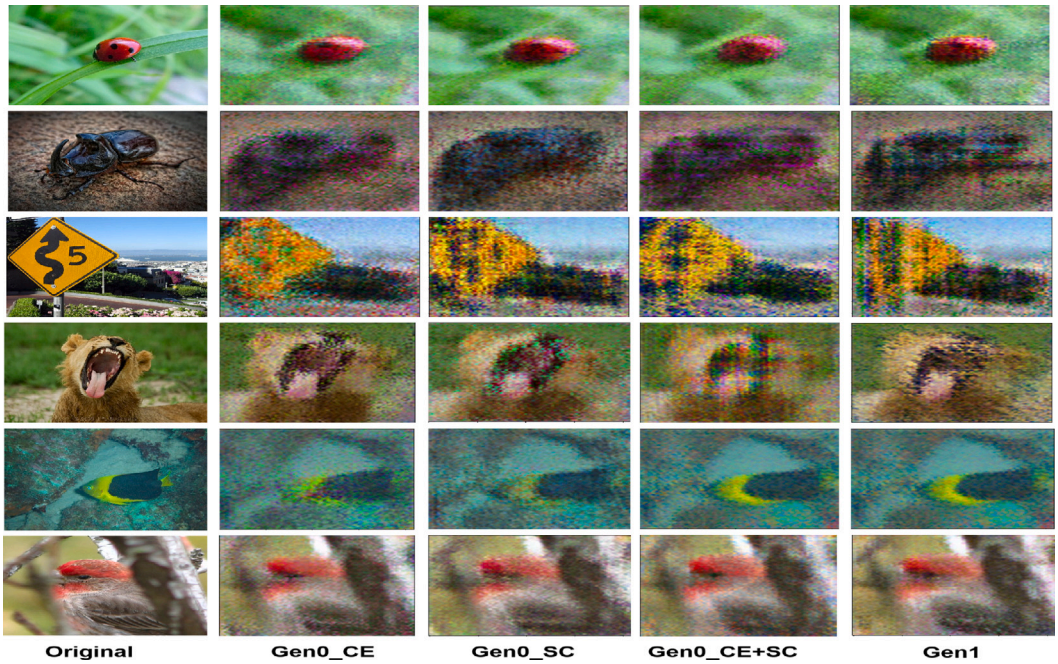


Fig. 7. Image reconstruction of features extracted by each component of the proposed method.

Table 4
Comparison of test accuracy (%) with related methods under 5-way 1-shot and 5-shot tasks on miniImageNet.

Method	Backbone	Empirical error		Regularization					Performance	
		CE	SC	MM	NM	SSL	SD	SLA	1-shot	5-shot
Baseline1	ResNet12	✓	×	×	×	×	×	×	59.65 ± 0.45	79.57 ± 0.31
Baseline2	ResNet12	×	✓	×	×	×	×	×	62.34 ± 0.45	75.76 ± 0.35
Neg-Cosine [19]	WRN28	✓	×	×	✓	×	×	×	61.72 ± 0.81	81.79 ± 0.55
S2M2 [18]	WRN28	✓	×	✓	×	✓	×	×	64.93 ± 0.18	83.18 ± 0.11
RFS [17]	WRN28	✓	×	×	×	×	✓	×	64.82 ± 0.60	82.14 ± 0.43
SKD [48]	ResNet12	✓	×	×	×	✓	✓	×	67.04 ± 0.85	83.54 ± 0.54
CSIV [9]	ResNet12	✓	×	×	×	✓	✓	×	67.28 ± 0.80	84.78 ± 0.33
PAL [49]	ResNet12	✓	✓	×	×	✓	✓	×	69.37 ± 0.64	84.40 ± 0.44
BANs_SLA	ResNet12	✓	✓	×	×	×	✓	✓	**70.40 ± 0.44**	**85.31 ± 0.22**

'✓' means this term is used in the method. '×' means that this term is not used in the method.

Table 5
Comparison of results on miniImageNet, tiredImageNet and CIFAR-FS.

Method	Backbone	Venue	miniImageNet		tiredImageNet		CIFAR-FS	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Meta-learning								
Prototypical[a] [7]	Conv4	NIPS'17	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74	–	–
Relational[a] [8]	Conv4	CVPR'18	50.44 ± 0.82	65.32 ± 0.70	54.48 ± 0.93	71.32 ± 0.78	55.00 ± 1.00	69.30 ± 0.80
DeepEMD [34]	ResNet12	CVPR'20	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58	–	–
CC+rot [45]	ResNet12	CVPR'20	62.93 ± 0.45	79.87 ± 0.33	70.53 ± 0.51	84.98 ± 0.36	76.09 ± 0.30	87.83 ± 0.21
BML [50]	ResNet12	ICCV'21	67.04 ± 0.63	83.63 ± 0.29	68.99 ± 0.50	85.49 ± 0.34	73.45 ± 0.47	88.04 ± 0.33
RENet [31]	ResNet12	ICCV'21	67.60 ± 0.44	82.58 ± 0.30	71.61 ± 0.51	85.28 ± 0.35	74.51 ± 0.46	86.60 ± 0.32
MeTAL [29]	ResNet12	CVPR'21	66.61 ± 0.28	81.43 ± 0.25	70.29 ± 0.40	86.17 ± 0.35	–	–
DAN [35]	ResNet12	CVPR'21	67.76 ± 0.46	82.71 ± 0.31	71.89 ± 0.52	85.96 ± 0.35	–	–
IEPT [44]	ResNet12	ICLR'21	67.05 ± 0.44	82.90 ± 0.30	72.24 ± 0.50	86.73 ± 0.34	–	–
APP2S [53]	ResNet12	AAAI'22	66.25 ± 0.20	83.42 ± 0.15	72.00 ± 0.22	86.23 ± 0.15	73.12 ± 0.22	85.69 ± 0.16
DeepBDC [39]	ResNet12	CVPR'22	67.34 ± 0.43	84.46 ± 0.28	72.34 ± 0.49	87.31 ± 0.32	–	–
MFS [32]	ResNet12	CVPR'22	68.32 ± 0.62	82.71 ± 0.46	73.63 ± 0.88	87.59 ± 0.57	–	–
HGNN [36]	ResNet12	AAAI'22	67.02 ± 0.20	83.00 ± 0.13	72.05 ± 0.23	86.49 ± 0.15	–	–
Transfer learning								
Baseline++ [16]	ResNet12	ICLR'19	48.24 ± 0.75	66.43 ± 0.63	–	–	–	
Neg-Cosine [19]	WRN28	ECCV'20	61.72 ± 0.81	81.79 ± 0.55	–	–	–	
RFS [17]	WRN28	ECCV'20	64.82 ± 0.60	82.14 ± 0.43	71.52 ± 0.69	86.03 ± 0.49	–	–
CBM [54]	ResNet12	MM'20	64.77 ± 0.46	80.50 ± 0.33	71.27 ± 0.50	85.81 ± 0.34	–	–
SKD [48]	ResNet12	Arxiv'21	67.04 ± 0.85	83.54 ± 0.54	72.03 ± 0.91	86.50 ± 0.58	76.90 ± 0.9	88.9 ± 0.60
CSIV [9]	ResNet12	CVPR'21	67.28 ± 0.80	84.78 ± 0.33	72.21 ± 0.90	87.08 ± 0.58	77.87 ± 0.85	89.74 ± 0.57
PAL [49]	ResNet12	ICCV'21	69.37 ± 0.64	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47	77.10 ± 0.70	88.0 ± 0.50
CCF [38]	ResNet12	CVPR'22	68.88 ± 0.43	84.59 ± 0.30	–	–	–	–
GLFA [55]	ResNet12	PR'23	67.25 ± 0.36	82.80 ± 0.30	72.25 ± 0.40	86.37 ± 0.27	–	–
BANs_SLA	ResNet12	–	**70.40 ± 0.44**	**85.31 ± 0.22**	**73.85 ± 0.49**	**87.72 ± 0.33**	**77.98 ± 0.45**	**89.78 ± 0.31**

'–' Means the results are not provided by the authors. The best results are in bold font.
[a]Means the results are reported in [34].

space learned by Gen1 than those learned by Gen0_CE, Gen0_SC, and Gen0_CE+SC. The samples in the same class gather tighter with clear boundaries away from different classes, indicating that our method can obtain discriminative and robust feature representations.

This robustness is key for handling the diversity of novel class features, ensuring that the model can generalize well to unseen data. This analysis validates our theoretical discussion, demonstrating that our method effectively minimizes the generalization error on the base set, enhancing its generalization ability on unseen data.

4.8. Image reconstruction of features

Our method follows transfer learning, which emphasizes learning good feature representation during the pre-training. In order to intuitively show the effectiveness of our method, we display what features have been retained by the pre-trained backbone network of each component in our method. We use deep image prior [42] to invert the features extracted by the pre-trained backbone network into RGB images. The reconstruction results are shown in Fig. 7. We notice that CE loss allows for the holistic information of objects, while SC loss highlights the details and contours of objects. Combining them together can keep their advantages, and then reconstructed images appear good global state with details. Finally, our method can well reconstruct the original image due to the instruction of a pre-trained teacher.

4.9. Comparison with most related methods

Our method is most related to the methods leveraging various loss functions in the context of transfer learning, including Neg-Cosine [19], RFS [17], S2M2 [18], SKD [48], PAL [49]. The comparison of results between our method and these methods on the most popular dataset of miniImageNet are shown in Table 4. These loss functions are broadly categorized into two groups: empirical error-based and regularization-based. The typical ones of the former are CE and SC losses. The latter appeared in the area of FSC can be summarized as Negative Margin (NM), Manifold Mixup (MM), Self-Supervised Learning (SSL), Self-Distillation (SD) and Self-Supervised Label Augmentation (SLA). Please note that we re-implement Baseline1 and Baseline2 with our own code, which respectively only uses CE and SC on the original base dataset. From the results, it is observed that: (1) Compared with Baseline1, all the regularization techniques can improve the FSC performance. This illustrates that current methods have already followed our proposed theorem, which summarizes them and forms a complete edition to well explain the process of pre-training. (2) Of all the comparison results, PAL and CSIV have outstanding performance. However, they require many loss terms in the final optimization function. Our method is about 1.03% and 0.53% higher than them respectively in 1-shot and 5-shot classification, but only has one tuned parameter due to an implicit regularization framework of SLA. Our method is simple yet effective and the results further validated our proposed theorem experimentally.

Table 6
Comparison of results on CUB dataset.

Method	CUB	
	1-shot	5-shot
Meta-learning		
DeepEMD [34]	75.65 ± 0.83	88.69 ± 0.50
BML [50]	76.21 ± 0.63	90.45 ± 0.36
RENet [31]	79.49 ± 0.44	91.11 ± 0.24
IEPT [44]	69.97 ± 0.49	84.33 ± 0.33
APP2S [53]	77.64 ± 0.19	90.43 ± 0.18
MFS [32]	79.60 ± 0.80	90.48 ± 0.44
HGNN [36]	78.58 ± 0.20	90.02 ± 0.12
Transfer learning		
Baseline++ [16]	60.53 ± 0.83	79.34 ± 0.61
Neg-Cosine [19]	72.66 ± 0.85	89.40 ± 0.43
S2M2 [18]	80.68 ± 0.81	90.85 ± 0.44
CCF [38]	81.85 ± 0.42	91.58 ± 0.32
GLFA [55]	76.52 ± 0.37	90.27 ± 0.38
BANs_SLA	**83.17 ± 0.39**	**93.75 ± 0.19**

The best results are in bold font.

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

We compare the performance of BANs_SLA with several State-of-the-Art (SOTA) methods on four popular datasets. According to the learning paradigm, these methods are broadly classified into two categories: meta-learning based and transfer learning based. From the comparison results shown in Tables 5 and 6, we can see that: (1) Compared with meta-learning methods, our method has achieved better performance. Specifically, on miniImageNet, MFS and DeepBDC behave the best in 1-shot and 5-shot settings, respectively. our method beats them 2.08% and 0.85%, respectively. On tiredImageNet, our method outperforms the best MFS by 0.22% and 0.13% respectively in 1-shot and 5-shot settings. On CIFAR-FS, our method achieves 3.38% and 1.74% improvements than RENet and BML for 1-shot and 5-shot respectively. On CUB, our method exceeds the best MFS by 3.57% and 3.27% respectively in 1-shot and 5-shot settings. (2) Compared with transfer learning methods, our method also has shown better performance. Specifically, on miniImageNet, PAL and CSIV behave the best respectively in 1-shot and 5-shot settings, our method beats them by 1.03% and 0.53%. On tiredImageNet, our method outperforms the best PAL and CSIV by 1.60% and 0.64% respectively in 1-shot and 5-shot settings. On CIFAR-FS, our method achieves 0.11% and 0.04% improvement over CSIV for 1-shot and 5-shot respectively. On CUB, our method exceeds the best CCF by 1.32% and 2.17% respectively in 1-shot and 5-shot settings. In a word, our method consistently outperforms current state-of-the-art FSC methods under both 1-shot and 5-shot tasks on multiple datasets. The promising performance is mainly attributed to minimizing the empirical error and adopting effective regularization strategies simultaneously.

5. Conclusions

This paper proposes the generalization error bound theorem as the general rule to guide the FSC learning process in the context of transfer learning. From this theorem, we learn that the pre-training stage shall aim at minimizing the base-class generalization error. Following this idea, we design a method called Born-Again Networks under Self-supervised Label Augmentation (BANs-SLA) to decrease the base-class generalization error by investigating the empirical error and regularization techniques jointly. Extensive results have validated the effectiveness of each component in BANs-SLA, which have supported our theorem. Moreover, just as stated in Theorem 1, the classification error bound on novel classes is mainly determined three terms of (1)base-class generalization error, (2) the base-novel domain divergence and (3) the novel-class generalization error produced by an incremental learner using novel samples. Our method mainly focuses on the first term, limiting in addressing the other two terms. Especially, the second term of domain divergence has significant impact on the FSC performance. In our future work, we plan to explore techniques such as adversarial domain adaptation to mitigate the domain divergence issue.

Declaration of competing interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability

Data will be made available on request.

Appendix. Supplementary material

A.1. Proof of Theorem 1

Proof.

$$\begin{aligned} e_n(h^o) &\approx e_n(h) + e_n(h^\triangle) \\ &= e_n(h) + e_b(h) - e_b(h) + e_{D_b}(h, f_n) - e_{D_b}(h, f_n) + e_n(h^\triangle) \\ &= e_b(h) + e_{D_b}(h, f_n) - e_b(h) + e_n(h) - e_{D_b}(h, f_n) + e_n(h^\triangle). \end{aligned} \tag{13}$$

Substitute Eq. (2) into the right side of Eq. (13), then:

$$\begin{aligned} & e_n(h^o) = e_b(h) + e_{D_b}(h, f_n) - e_{D_b}(h, f_b) + e_{D_n}(h, f_n) - \\ & e_{D_b}(h, f_n) + e_n(h^\triangle) \\ & \le e_b(h) + \left| e_{D_b}(h, f_n) - e_{D_b}(h, f_b) \right| + \\ & \left| e_{D_n}(h, f_n) - e_{D_b}(h, f_n) \right| + e_n(h^\triangle) \\ & \le e_b(h) + \left| E_{X \in D_b}[\,|h(x) - f_n(x)|\,] - E_{X \in D_b}[\,|h(x) - f_b(x)|\,] \right| \\ & + \left| e_{D_n}(h, f_n) - e_{D_b}(h, f_n) \right| + e_n(h^\triangle). \end{aligned} \tag{14}$$

As the expected absolute value is less than or equal to the expectation of the absolute value, the In Eq. (14) becomes:

$$\begin{aligned} & e_n(h^o) \le e_b(h) + E_{X \in D_b}[\,|h(x) - f_n(x)| - |h(x) - f_b(x)|\,] + \\ & \left| e_{D_n}(h, f_n) - e_{D_b}(h, f_n) \right| + e_n(h^\triangle). \end{aligned} \tag{15}$$

For $h(x) \in [0, 1]$ and $f(x) \in [0, 1]$, then:

$$\begin{aligned} & e_n(h^o) \le e_b(h) + E_{X \in D_b} \left| f_n(x) - f_b(x) \right| \\ & + \left| e_{D_n}(h, f_n) - e_{D_b}(h, f_n) \right| + e_n(h^\triangle). \end{aligned} \tag{16}$$

Let $E_{X \in D_b} \left| f_n(x) - f_b(x) \right|$ be λ, the In Eq. (16) becomes:

$$\begin{aligned} & e_n(h^o) \le e_b(h) + \lambda + \left| e_{D_n}(h, f_n) - e_{D_b}(h, f_n) \right| \\ & + e_n(h^\triangle). \end{aligned} \tag{17}$$

Let the samples in the base domain and the novel domain construct the whole sample space of $\mathcal{B}$, the probability distribution of two domains are respectively denoted as $\phi_b(x)$ and $\phi_n(x)$, then the In Eq. (17) becomes:

$$\begin{aligned} & e_n(h^o) \le e_b(h) + \lambda + \left| \int_{\mathcal{B}} \left| h(x) - f_n(x) \right| (\phi_b(x) - \phi_n(x)) dx \right| \\ & + e_n(h^\triangle). \end{aligned} \tag{18}$$

$\mathcal{B}$ is subdivided into two spaces $\mathcal{B}_1$ and $\mathcal{B}_2$, then:

$$\begin{aligned} & e_n(h^o) \le e_b(h) + \lambda + \left| \int_{\mathcal{B}_1} \left| h(x) - f_n(x) \right| (\phi_b(x) - \phi_n(x)) dx \right| \\ & + \left| \int_{\mathcal{B}_2} \left| h(x) - f_n(x) \right| (\phi_b(x) - \phi_n(x)) dx \right| + e_n(h^\triangle). \end{aligned} \tag{19}$$

Since $|h(x) - f_n(x)| \leq 1$, then:

$$\begin{aligned} e_n(h^o) &\leq e_b(h) + \lambda + \left| \int_{\mathcal{X}_1} (\phi_b(x) - \phi_n(x)) dx \right| + \\ &\left| \int_{\mathcal{X}_2} (\phi_b(x) - \phi_n(x)) dx \right| + e_n(h^\triangle) \\ &\leq e_b(h) + \lambda + \left| Pr(\mathcal{B}_1)_{D_b} - Pr(\mathcal{B}_1)_{D_n} \right| + \\ &\left| Pr(\mathcal{B}_2)_{D_b} - Pr(\mathcal{B}_2)_{D_n} \right| + e_n(h^\triangle) \\ &\leq e_b(h) + \lambda + 2 \sup_{B \in \mathcal{B}} \left| Pr_{D_b}(B) - Pr_{D_n}(B) \right| + e_n(h^\triangle). \end{aligned} \tag{20}$$

Substitute Eq. (3) into the right side of In Eq. (20), we get:

$$e_n(h^o) \leq e_b(h) + \lambda + \mathcal{L}(D_b, D_n) + e_n(h^\triangle). \tag{21}$$

Furthermore, according to the statistical learning theory [21], for every $h \in H$, the relationship between the empirical risk $\hat{e}_b(h)$ and the expected risk $e_b(h)$on D_b with probability at least 1-η has:

$$e_b(h) \leq \hat{e}_b(h) + \sqrt{\frac{v(ln\frac{2L}{v} + 1) - ln\frac{\eta}{4}}{L}}. \tag{22}$$

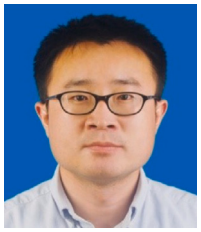
With the same theory, we get the relationship between the empirical risk $\hat{e}_n(h^\triangle)$ and the expected risk $e_n(h^\triangle)$ on D_n:

$$e_n(h^\triangle) \leq \hat{e}_n(h^\triangle) + \sqrt{\frac{v(ln\frac{2K}{v} + 1) - ln\frac{\eta}{4}}{K}}. \tag{23}$$

Substitute the In Eq. (22) and (23) into the right side of In Eq. (21), we finally get the few-shot generalization error bound theorem.

References

[1] L. Fe-Fei, et al., A Bayesian approach to unsupervised one-shot learning of object categories, in: ICCV, 2003, pp. 1134–1141.
[2] B.M. Lake, R. Salakhutdinov, J.B. Tenenbaum, Human-level concept learning through probabilistic program induction, Science 350 (6266) (2015) 1332–1338.
[3] P. Bateni, R. Goyal, V. Masrani, F. Wood, L. Sigal, Improved few-shot visual classification, in: CVPR, 2020, pp. 14493–14502.
[4] Q. Fan, W. Zhuo, C.-K. Tang, Y.-W. Tai, Few-shot object detection with attention-RPN and multi-relation detector, in: CVPR, 2020, pp. 4013–4022.
[5] W. Liu, C. Zhang, G. Lin, F. Liu, Crnet: Cross-reference networks for few-shot segmentation, in: CVPR, 2020, pp. 4165–4173.
[6] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: ICML, 2017, pp. 1126–1135.
[7] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: NIPS, 2017.
[8] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: relation network for few-shot learning, in: CVPR, 2018, pp. 1199–1208.
[9] M.N. Rizve, S. Khan, F.S. Khan, M. Shah, Exploring complementary strengths of invariant and equivariant representations for few-shot learning, in: CVPR.
[10] X. Li, X. Yang, Z. Ma, J.-H. Xue, Deep metric learning for few-shot image classification: A Review of recent developments, Pattern Recognit. (2023) 109381.
[11] L. Collins, A. Mokhtari, S. Shakkottai, Task-robust model-agnostic meta-learning, in: NIPS, 2020, pp. 18860–18871.
[12] S. Baik, S. Hong, K.M. Lee, Learning to forget for meta-learning, in: CVPR, 2020, pp. 2379–2387.
[13] C. Simon, P. Koniusz, R. Nock, M. Harandi, Adaptive subspaces for few-shot learning, in: CVPR, 2020, pp. 4136–4145.
[14] B. Zhang, X. Li, Y. Ye, Z. Huang, L. Zhang, Prototype completion with primitive knowledge for few-shot learning, in: CVPR, 2021, pp. 3754–3762.
[15] Q. Liu, W. Cao, Z. He, Cycle optimization metric learning for few-shot classification, Pattern Recognit. 139 (2023) 109468.
[16] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C.F. Wang, J.-B. Huang, A closer look at few-shot classification, in: ICLR, 2019.
[17] Y. Tian, Y. Wang, D. Krishnan, J.B. Tenenbaum, P. Isola, Rethinking few-shot image classification: a good embedding is all you need? in: ECCV, 2020, pp. 266–282.
[18] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, V.N. Balasubramanian, Charting the right manifold: Manifold mixup for few-shot learning, in: CVPR, 2020, pp. 2218–2227.
[19] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, H. Hu, Negative margin matters: Understanding margin in few-shot classification, in: ECCV, 2020, pp. 438–455.
[20] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach. Learn. 79 (1) (2010) 151–175.
[21] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (5) (1999) 988–999.
[22] H. Lee, S.J. Hwang, J. Shin, Self-supervised label augmentation via input transformations, in: ICML, 2020, pp. 5714–5724.
[23] L. Wei, L. Xie, J. He, J. Chang, X. Zhang, W. Zhou, H. Li, Q. Tian, Can semantic labels assist self-supervised visual representation learning? 2020, arXiv preprint arXiv:2011.08621.
[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: NIPS, 2020, pp. 18661–18673.
[25] H. Mobahi, M. Farajtabar, P. Bartlett, Self-distillation amplifies regularization in hilbert space, in: NIPS, 2020, pp. 3351–3361.
[26] A. Nichol, J. Schulman, Reptile: a scalable metalearning algorithm, 2018, arXiv preprint arXiv:1803.02999.
[27] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: ICLR, 2016.
[28] L. Bertinetto, J.F. Henriques, P.H. Torr, A. Vedaldi, Meta-learning with differentiable closed-form solvers, in: ICML, 2019.
[29] S. Baik, J. Choi, H. Kim, D. Cho, J. Min, K.M. Lee, Meta-learning with task-adaptive loss function for few-shot learning, in: CVPR, 2021, pp. 9465–9474.
[30] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one-shot learning, in: NIPS, 2016.
[31] D. Kang, H. Kwon, J. Min, M. Cho, Relational embedding for few-shot classification, in: CVPR, 2021, pp. 8822–8833.
[32] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, C. Gagné, Matching feature sets for few-shot image classification, in: CVPR, 2022, pp. 9014–9024.
[33] J. Wu, T. Zhang, Y. Zhang, F. Wu, Task-aware part mining network for few-shot learning, in: CVPR, 2021, pp. 8433–8442.
[34] C. Zhang, Y. Cai, G. Lin, C. Shen, DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers, in: CVPR, 2020, pp. 12203–12213.
[35] C. Xu, Y. Fu, C. Liu, C. Wang, J. Li, F. Huang, L. Zhang, X. Xue, Learning dynamic alignment via meta-filter for few-shot learning, in: CVPR, 2021, pp. 5182–5191.
[36] T. Yu, S. He, Y.-Z. Song, T. Xiang, Hybrid graph neural networks for few-shot learning, in: AAAI, 2022, pp. 3179–3187.
[37] S. Yang, L. Liu, M. Xu, Free lunch for few-shot learning: distribution calibration, in: ICLR, 2021.
[38] J. Xu, X. Pan, X. Luo, W. Pei, Z. Xu, Exploring category-correlated feature for few-shot image classification, 2021, arXiv preprint arXiv:2112.07224.
[39] J. Xie, F. Long, J. Lv, Q. Wang, P. Li, Joint distribution matters: deep Brownian distance covariance for few-shot classification, in: CVPR, 2022, pp. 7972–7981.
[40] Z. Zhao, Q. Liu, W. Cao, D. Lian, Z. He, Self-guided information for few-shot classification, Pattern Recognit. 131 (2022) 108880.
[41] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: CVPR, 2020, pp. 9729–9738.
[42] N. Zhao, Z. Wu, R.W. Lau, S. Lin, What makes instance discrimination good for transfer learning? in: ICLR, 2021.
[43] A. Islam, C.-F.R. Chen, R. Panda, L. Karlinsky, R. Radke, R. Feris, A broad study on the transferability of visual representations with contrastive learning, in: CVPR, 2021, pp. 8845–8855.
[44] M. Zhang, J. Zhang, Z. Lu, T. Xiang, M. Ding, S. Huang, IEPT: Instance-level and episode-level pretext tasks for few-shot learning, in: ICLR, 2020.
[45] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, M. Cord, Boosting few-shot visual learning with self-supervision, in: CVPR, 2019, pp. 8059–8068.
[46] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, A. Anandkumar, Born again neural networks, in: ICML, 2018, pp. 1607–1616.
[47] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: CVPR, 2018, pp. 4320–4328.
[48] J. Rajasegaran, S. Khan, M. Hayat, F.S. Khan, M. Shah, Self-supervised knowledge distillation for few-shot learning, 2020, arXiv preprint arXiv:2006.09785.
[49] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, W. Abd-Almageed, Partner-assisted learning for few-shot image classification, in: CVPR, 2021, pp. 10573–10582.
[50] Z. Zhou, X. Qiu, J. Xie, J. Wu, C. Zhang, Binocular mutual learning for improving few-shot classification, in: CVPR, 2021, pp. 8402–8411.
[51] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J.B. Tenenbaum, H. Larochelle, R.S. Zemel, Meta-learning for semi-supervised few-shot classification, 2018, arXiv preprint arXiv:1803.00676.
[52] N. Hilliard, L. Phillips, S. Howland, A. Yankov, C.D. Corley, N.O. Hodas, Few-shot learning with metric-agnostic conditional embeddings, 2018, arXiv preprint arXiv:1802.04376.
[53] R. Ma, P. Fang, T. Drummond, M. Harandi, Adaptive poincaré point to set distance for few-shot classification, in: AAAI, 2022, pp. 1926–1934.
[54] Z. Wang, Y. Zhao, J. Li, Y. Tian, Cooperative bi-path metric for few-shot learning, in: ACM MM, 2020, pp. 1524–1532.
[55] B. Shi, W. Li, J. Huo, P. Zhu, L. Wang, Y. Gao, Global-and local-aware feature augmentation with semantic orthogonality for few-shot image classification, Pattern Recognit. 139 (2023).



Fan Liu is currently a professor of Hohai University. He received his B.S. degree and Ph.D. degree from Nanjing University of Science and Technology (NUST) in 2009 and 2015. From September 2008 to December 2008, he studied at Ajou University in South Korea. From February 2014 to May 2014, he worked at Microsoft Research Asia. His research interests include computer vision, pattern recognition, and machine learning. Dr. Liu serves as a reviewer of IEEE TNNLS, IEEE TKDE, ACM TIST, Information Sciences, Neurocomputing, Pattern Analysis and Application and an executive director of Jiangsu association of Artificial Intelligence (JSAI).

Sai Yang is currently a lecturer of Nantong University. She received her M.S. degree from School of Mechanical and Electrical Engineering, Jiangxi University of Science and Technology, China, in 2010, and a Ph.D. degree from School of Computer Science and Engineering, Nanjing University of Science and Technology, China, in 2015. Her research interests include computer vision, image processing, pattern recognition and machine learning.

Delong Chen received the B.Sc. degree of computer science from Hohai University, Nanjing, China in 2021. He is currently a research intern at XiaobingAI. His research interest includes vision-language, multi-modal large language models, and computer vision.

Huaxi Huang received the B.Eng. degree and the M. Eng. degree in computer science from Tianjin University, China, in 2014 and 2017, respectively. He also obtained his Ph.D. degree in Data Analytics from the University of Technology Sydney, Australia, in 2022. Currently, he is a CERC Postdoctoral Research Fellow with Data61 at the Commonwealth Scientific and Industrial Research Organisation in Australia. His research interests include multimedia analysis, computer vision, and machine learning.

Jun Zhou (Senior Member, IEEE) received the B.S.degree in computer science and the B.E. degree in international business from the Nanjing University of Science and Technology, Nanjing, China, in 1996 and 1998, respectively, the M.S. degree in computer science from Concordia University, Montreal, QC, Canada, in 2002, and the Ph.D. degree in computing science from the University of Alberta, Edmonton, AB, Canada, in 2006.

He was a Research Fellow with the Research School of Computer Science, The Australian National University, Canberra, ACT, Australia, and a Researcher with the Canberra Research Laboratory, National Information and Communications Technology Australia, Canberra, Australia. In 2012, he joined the School of Information and Communication Technology, Griffith University, Nathan, QLD, Australia, where he is an Associate Professor . His research interests include pattern recognition, computer vision, and spectral imaging and their applications in remote sensing and environmental informatics.