



Hybrid unsupervised representation learning and pseudo-label supervised self-distillation for rare disease imaging phenotype classification with dispersion-aware imbalance correction

Jinghan Sun ^{a,b,1,2}, Dong Wei ^{b,1}, Liansheng Wang ^{a,*}, Yefeng Zheng ^b

^a Xiamen University, Xiamen, 361005, China

^b Jarvis Research Center, Tencent YouTu Lab, Shenzhen, 518000, China

ARTICLE INFO

Keywords:

Rare disease classification
Unsupervised representation learning
Pseudo-label supervised self-distillation
Dispersion-aware imbalance correction

ABSTRACT

Rare diseases are characterized by low prevalence and are often chronically debilitating or life-threatening. Imaging phenotype classification of rare diseases is challenging due to the severe shortage of training examples. Few-shot learning (FSL) methods tackle this challenge by extracting generalizable prior knowledge from a large base dataset of common diseases and normal controls and transferring the knowledge to rare diseases. Yet, most existing methods require the base dataset to be labeled and do not make full use of the precious examples of rare diseases. In addition, the extremely small size of the training samples may result in inter-class performance imbalance due to insufficient sampling of the true distributions. To this end, we propose in this work a novel hybrid approach to rare disease imaging phenotype classification, featuring three key novelties targeted at the above drawbacks. First, we adopt the unsupervised representation learning (URL) based on self-supervising contrastive loss, whereby to eliminate the overhead in labeling the base dataset. Second, we integrate the URL with pseudo-label supervised classification for effective self-distillation of the knowledge about the rare diseases, composing a hybrid approach taking advantage of both unsupervised and (pseudo-) supervised learning on the base dataset. Third, we use the feature dispersion to assess the intra-class diversity of training samples, to alleviate the inter-class performance imbalance via dispersion-aware correction. Experimental results of imaging phenotype classification of both simulated (skin lesions and cervical smears) and real clinical rare diseases (retinal diseases) show that our hybrid approach substantially outperforms existing FSL methods (including those using a fully supervised base dataset) via effective integration of the URL, pseudo-label driven self-distillation, and dispersion-aware imbalance correction, thus establishing a new state of the art.

1. Introduction

Rare diseases are a significant public health issue and a challenge to healthcare. On aggregate, the number of people suffering from rare diseases worldwide is estimated over 400 million, and there are about 5000–7000 rare diseases—with 250 new ones appearing each year (Stolk et al., 2006). Patients with rare diseases face delayed diagnosis: 10% of patients spent 5–30 years to reach a final diagnosis. Besides, many rare diseases can be misdiagnosed. Therefore, accurate imaging phenotype classification to facilitate the timely diagnosis of rare diseases can be of great clinical value. In recent years, deep learning (DL) methods have developed into the state of the art (SOTA) for image-based computer-aided diagnosis (CAD) of many diseases (Ker et al., 2018; Litjens et al., 2017; Shen et al., 2017). However, due to

the limited number of patients for a specific rare disease, collecting sufficient data for well training of generic DL classification models can be practically difficult or even infeasible for rare diseases.

To cope with the scarcity of training samples, a machine learning paradigm called few-shot learning (FSL) has been proposed (Li et al., 2006) and achieved remarkable advances in the natural image domain (Finn et al., 2017; Hsu et al., 2018; Khodadadeh et al., 2019; Shi et al., 2022; Snell et al., 2017; Vinyals et al., 2016). In FSL, generalizable prior knowledge is learned on a large dataset of base classes, and subsequently utilized to boost learning of previously unseen novel classes given limited samples (the target task). Earlier approaches (Finn et al., 2017; Hsu et al., 2018; Khodadadeh et al., 2019; Snell et al., 2017; Vinyals et al., 2016) to FSL mostly resorted to the concept of

* Corresponding author.

E-mail addresses: lswang@xmu.edu.cn (L. Wang), yefengzheng@tencent.com (Y. Zheng).

¹ Contributed equally.

² J. Sun contributed to this work during an internship at Tencent.

meta-learning and involved complicated framework design and task construction. Recently, Tian et al. (2020) showed that superior FSL performance could be achieved by simply learning a good representation on the base dataset using basic frameworks, followed by fitting a simple classifier to a few examples of the novel classes. Additional performance boosts were achieved through self-distillation (Furlanello et al., 2018; Hinton et al., 2015). How to implement an effective representation learning plus self-distilling strategy on an unsupervised base dataset, though, is not obvious.

As to FSL for medical image classification, we are aware of only a few existing works (Chen et al., 2022; Jiang et al., 2019; Li et al., 2020; Paul et al., 2021; Zhu et al., 2020), and to the best of our knowledge, all of them relied on heavy labeling of the base dataset, causing a great burden for practical applications. Moreover, the meta-learning process and the target task were often isolated in most existing FSL approaches, and the meta-learner had little knowledge about its end task. For natural images, this setting is consistent with the general purpose of pretraining a classifier that can be quickly adapted for diverse tasks. For the scenario we consider, however, known types of rare diseases are mostly fixed, and their recognition constitutes a definite task. We hypothesize that, by bridging the base dataset and the definite task, the performance can be boosted for the rare disease classification.

In addition, most existing works only focused on the overall model performance (e.g., total accuracy), but ignored the inter-class performance gaps. Due to the extremely small size, the available scarce training data are highly sensitive to the randomness in sampling and often cannot fully represent each class. For example, the few data points of a specific class may be sufficiently diverse while those of another may be quite similar to each other, which may lead to good performance on the former but poor on the latter. This performance imbalance between classes looks similar to the imbalance problem frequently encountered in long-tail distributed classification, where the solutions include the classical re-sampling or re-weighting, and various more recently proposed training loss functions (Cao et al., 2019; Cui et al., 2019; Tan et al., 2021; Wang et al., 2021). However, as these solutions were designed to tackle the imbalanced size distribution of the classes, they were not applicable to our scenario in this work—where the number of samples for each class is the same in most FSL settings.

In this work, we propose a novel hybrid approach to rare disease imaging phenotype classification, which combines unsupervised representation learning (URL), pseudo-label supervised self-distillation (Furlanello et al., 2018; Hinton et al., 2015), and dispersion-aware imbalance correction. Motivated by the recent surge of representation learning in FSL (Chen et al., 2019; Tian et al., 2020), we first build a simple yet effective baseline model based on URL, where a good representation is learned on a large unlabeled base dataset consisting of common diseases and normal controls (CDNC) using contrastive learning (He et al., 2020), and applied to rare disease classification. So far as we are aware, this is the first study that explores few-shot medical image classification using an unsupervised base dataset. Then, we further propose to inject knowledge about the rare diseases into representation learning, to exploit the CDNC data for a more targeted learning of the rare diseases. Specifically, we use the baseline model as a teacher model to generate pseudo labels for instances of CDNC *belonging to the rare diseases*, to supervise knowledge distillation to the student model. Our rationale is that CDNC and rare diseases often share common characteristics, thus we can steer the representation learning on the former towards characteristics that can better distinguish the latter via supervision by the pseudo labels. In addition, we empirically explore design options for the distillation and find that a hybrid self-distillation integrating URL and pseudo-label supervised classification yields the best performance. Lastly, we introduce a dispersion-aware imbalance correction (DIC) strategy that takes into account intra-class feature dispersion to revise the model's prediction and reduce performance imbalance.

In summary, our contributions are three-fold:

- We propose a simple yet effective approach to rare disease imaging phenotype classification, which is based on URL and eliminates the need for labeling the large archive of CDNC data.
- We further propose to integrate the URL with pseudo-label supervised self-distillation, composing a hybrid approach taking advantage of both methodologies.
- We uncover the class imbalance issue in FSL for rare disease classification, and propose the DIC strategy to effectively mitigate the issue.

Thorough experiments on the ISIC 2018 skin lesion, Pap-smear cervical smear classification, and optical coherence tomography (OCT) diagnosis of rare retinal diseases datasets show that the URL-based baseline model already outperforms previous SOTA FSL methods (including those using a fully supervised base dataset), and that further boosts in performance are achieved with the proposed hybrid distillation and DIC.

This work is a comprehensive extension of our preliminary exploration (Sun et al., 2021) in three main aspects. First, by examining the per-class performance, we identify the class imbalance issue in FSL of rare diseases, which is practically relevant yet has been largely omitted in previous works. Second, methodologically, we extend our framework to address the imbalance issue with the newly proposed dispersion-aware imbalance correction on pseudo-labeled data. Third, we generalize our framework on two more public datasets and demonstrate consistently remarkable performance on few-shot rare disease imaging-phenotype classification.

2. Related works

2.1. Few-shot learning

Significant efforts have been devoted to few-shot learning (FSL) in the natural image domain, which can be organized into three main categories: metric learning, meta-learning, and representation learning. The core idea of metric learning is to learn a (set of) projection function(s) such that when projected in the embedding space, images can be easily classified based on certain distances. Vinyals et al. (2016) proposed to sample mini-batches called episodes during training, where each episode was designed to mimic the target task by subsampling classes as well as data points. They also compared cosine distance between embeddings of few labeled samples and query samples and employed a binary long short-term memory network (Hochreiter and Schmidhuber, 1997) to explore attention for labeled samples. Snell et al. (2017) inherited the episodic training scheme and used the Euclidean distance to learn the prototype center of each class. Instead of using a fixed metric function, Sung et al. (2018) adopted a relation network to learn the distance metric. However, these metric learning methods still faced the overfitting issue due to the scarcity of labeled data. To tackle this problem, meta-learning was proposed to limit the parameter space to avoid parameter overfitting in training set. MAML (Finn et al., 2017) and Reptile (Nichol and Schulman, 2018) provided a good parameter initialization to achieve rapid adaptation with limited novel training samples. However, these methods heavily relied on abundant labeled data of seen classes. Therefore, some works on unsupervised meta-learning focusing on the task construction mechanism without annotation were dedicated to alleviating this problem. CACTUs (Hsu et al., 2018) constructed episodes from unlabeled data via clustering, while incurring a heavy workload on the computational resources. Inspired by self-supervision, UMTRA (Khodadadeh et al., 2019) randomly labeled a subset of unlabeled data and applied online augmentation to them to construct episodes. Then, same as CACTUs, the constructed few-shot episodes can be used by any meta-learning framework.

Although meta-learning approaches have achieved decent performance on few-shot tasks, their framework design and task construction

procedure were complicated. In contrast, Tian et al. (2020) proposed a simple pipeline that only needed to learn a representation on the labeled base dataset, and then train a simple classifier on top of the representation using few samples of novel classes. The authors empirically proved that this simple non-episodic pipeline could achieve better performance. Some recent works (Boudiaf et al., 2021; Chen et al., 2019, 2021b) also analyzed the differences between the meta-learning algorithms and representation-based approaches, and demonstrated that a good representation could be more effective than complex meta-learning algorithms. Inspired by these work, we also propose to employ representation learning for rare disease classification, yet via contrastive learning on unsupervised base dataset to relieve the burden of manual annotation.

On the other hand, we are aware of only few works (Chen et al., 2022; Jiang et al., 2019; Li et al., 2020; Paul et al., 2021; Zhu et al., 2020; Zhang et al., 2020; Wu et al., 2023) on FSL of medical image classification, and to the best of our knowledge, a large annotation of the base dataset was required, causing a great burden for practical applications. In addition, most of the existing few-shot methods isolated the training process and target tasks, so the learned model had little knowledge about its end task. In this work, we adopt unsupervised representation learning on the base dataset, waiving the burden of labeling it. In addition, we propose to bridge the representation learning and the target task with a pseudo label supervised self-distillation, such that the learned representation is better purposed for its end task.

2.2. Contrastive learning

Contrastive learning can learn good representations from data without manual labels, where the contrastive loss (Hadsell et al., 2006) is employed to enforce features to be similar intra-class and dissimilar inter-classes. A large number of works (Chen et al., 2020a; Chen and He, 2021; Grill et al., 2020; He et al., 2020; Zbontar et al., 2021) defined various pretext tasks, such as applying various transformations to an image for maximizing mutual information between transformed image pairs. Contrastive learning can also be used for rare disease diagnosis to relieve the burden of manual labeling. Several works (Chen et al., 2021a; Liu et al., 2022; Shorfuazzaman and Hossain, 2021) proposed to learn feature representations by contrasting symptomatic and asymptomatic samples to reduce the requirement of large amounts of labeled data. However, they still relied on fine-tuning on large labeled datasets and could not easily adapt to few-shot tasks. Our method also incorporates contrastive learning to improve the network's representation ability on base classes, and further proposes self-distillation of rare disease knowledge given only few labeled samples of rare diseases.

2.3. Inter-class performance imbalance

The problem of inter-class performance imbalance is often encountered in learning with imbalanced classes. Re-sampling (Buda et al., 2018; Byrd and Lipton, 2019; Chawla et al., 2002) and re-weighting (Cao et al., 2019; Cui et al., 2019; Wang et al., 2021; Wei et al., 2021) are straightforward solutions that balance the numbers of training samples or weigh the importance of training losses according to the class sizes. Some works (Liu et al., 2020; Kim et al., 2020a; Ghiasi et al., 2021) also used data augmentation to supplement tail classes, sometimes directly in the latent space. Another option is to design specific new loss functions (Cao et al., 2019; Cui et al., 2019; Tan et al., 2021) for learning with imbalanced data. However, the performance imbalance in the context of few-shot rare disease classification does not originate from imbalanced classes, as the numbers of training samples in the support set are actually the same for each class. Therefore, the above-described methods may not be applicable. Alternatively, we resort to the feature dispersion (in contrast to the quantities of training samples) of the rare diseases estimated from pseudo-labeled data, for sensible adjustment and rebalance of the predicted class probabilities.

3. Methods

3.1. Background and problem setting

In the convention of the few-shot learning (FSL) literature (e.g., Vinyals et al., 2016), a *single* FSL classification task \mathcal{T} involves three datasets: a training set D , a support set S , and a query (testing) set Q . The support and query sets share the same label space and are used for training and testing, respectively. The target is to achieve optimal classification performance on Q . Assuming the support set S comprises K samples for each of N unique classes, where K is small, the FSL task is N -way K -shot. Note that the sizes of the support and query sets do not have to be the same (e.g., Snell et al., 2017). Although training a classifier solely on S is feasible, the performance is often unsatisfactory due to the small sample size. Hence, the large training set D —with a disjoint label space from the support and query sets—is exploited to learn a transferable representation to help construct a better classifier. A common meta-learning training strategy is called episodic training (Vinyals et al., 2016; Finn et al., 2017; Nichol and Schulman, 2018, to name a few), where each episode (a mini-batch indeed) is designed to mimic the FSL task by subsampling N classes and a pair of fake support and query sets \tilde{S} and \tilde{Q} from D to form a fake task $\tilde{\mathcal{T}}$, where $\tilde{\cdot}$ indicates a fake set/task with a disjoint label space from the actual target FSL task. The rationale is to make the training problem more faithful to the testing and thereby improve generalization. For performance evaluation, *multiple* FSL tasks are repeatedly constructed from another dataset disjoint from D , and the average performance of all tasks is reported.

Similar to Li et al. (2020), we formulate the task of rare disease imaging phenotype classification as an FSL problem. Concretely, we model a specific task as $\mathcal{T} = (S, Q)$ consisting of a support set $S = \{(x, y)\}$ and a query set $Q = \{(x, y)\}$, where x is an image and y its label. An N -way K -shot task includes N rare diseases, each with K instances in S , where K is small. Thus $y \in \{1, \dots, N\}$ and $|S| = N \times K$. To construct a task instance $\mathcal{T} \sim p(\mathcal{T})$, we randomly sample K examples for each of the N rare diseases from a rare disease dataset D_{rare} to form the support set S and use all remaining samples in D_{rare} to construct the query set Q . Only S is available for training and Q is solely for testing. Meanwhile, there is a large base dataset D_{base} consisting of common diseases and normal controls (CDNC). The target is to achieve optimal classification performance on Q given S and D_{base} . In this work, we consider D_{base} to be unlabeled for a more generally applicable approach in practice, by eliminating the need for annotation. Lastly, it is worth mentioning that a notable difference between our work and the existing meta-learning FSL approaches described in the previous paragraph is that, we follow Tian et al. (2020) to get rid of the episode training paradigm by randomly sampling mini-batches of training images from D_{base} . Therefore, no fake FSL learning task $\tilde{\mathcal{T}} = (\tilde{S}, \tilde{Q})$ is constructed during training.

3.2. Method overview

An overview of our method is shown in Fig. 1. Given D_{base} , we first perform unsupervised representation learning (URL) to train the embedding function f_q (Fig. 1(a)). Next, a simple classifier f_c is appended to the learned f_q (with parameters frozen) to compose a baseline model F (Fig. 1(b)), where f_c is optimized on S . Then, F is employed to assign each CDNC instance in D_{base} a pseudo label of the rare diseases (Fig. 1(c)). Lastly, a self-distillation via hybrid unsupervised and pseudo-label supervised representation learning with dispersion-aware imbalance correction (DIC) is performed on D_{base} to produce the final (student) model F' (Fig. 1(d)).

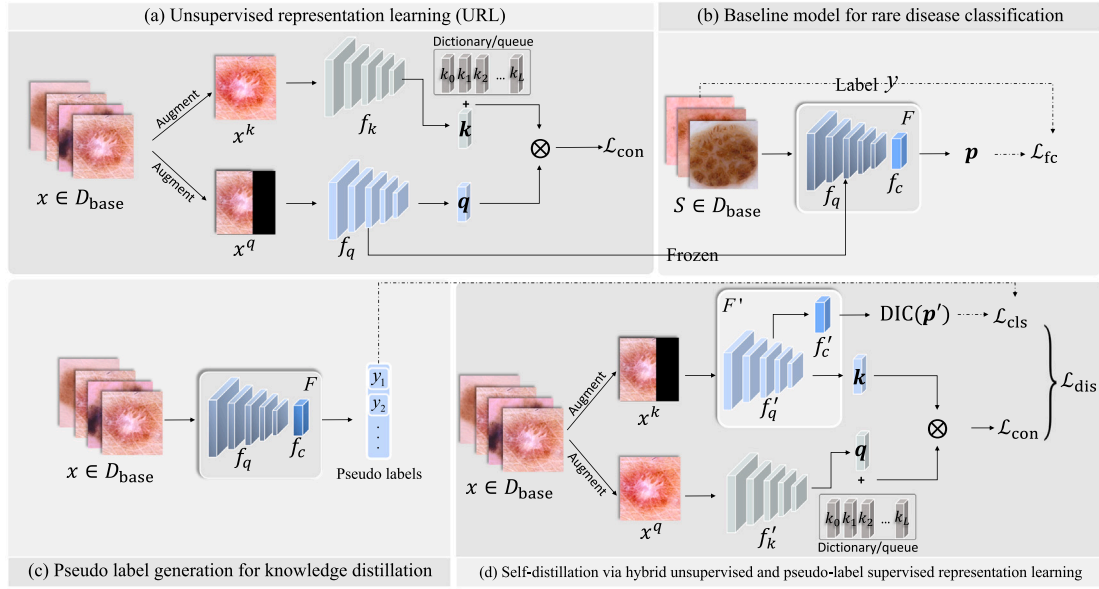


Fig. 1. Overview of the proposed approach. Solid line: information flow; dashed line: loss computation. Note that \mathcal{L}_{fc} in (b) can be any valid loss suitable for the classifier f_c . The dictionary maintains a queue of data samples' embedded representations. The current mini-batch is enqueued to the dictionary, and the oldest in the queue is removed.

3.3. URL on CDNC for rare disease classification

Inspired by the recent success of representation learning in FSL (Chen et al., 2019; Tian et al., 2020) and based on the recent advances in URL (Chen et al., 2020a; He et al., 2020), we propose to perform URL on the big yet unlabeled CDNC dataset for rare disease classification (Fig. 1(a)). Specifically, we adopt instance discrimination with the contrastive loss (He et al., 2020; Oord et al., 2018) as our self-supervising task. We employ MoCo_v1 (He et al., 2020), where each image x_i in D_{base} is augmented twice to obtain x_i^q and x_i^k , whose embedded representations are subsequently obtained by $q_i = f_q(x_i^q; \theta_q)$ and $k_i = f_k(x_i^k; \theta_k)$, where f_q and f_k are the query and key encoders parameterized by θ_q and θ_k , respectively. The contrastive loss \mathcal{L}_{con} is defined as (Oord et al., 2018):

$$\mathcal{L}_{\text{con}}(x_i) = -\log \left[\frac{\exp(q_i \cdot k_i / \tau)}{\exp(q_i \cdot k_i / \tau) + \sum_{j=1}^L \exp(q_i \cdot k_j / \tau)} \right], \quad (1)$$

where L is the number of keys stored in the dynamic dictionary implemented as a queue, and τ is a temperature hyperparameter. The dictionary maintains a queue of data samples' embedded representations. The current mini-batch is enqueued to the dictionary, and the oldest in the queue is removed. Intuitively, this loss is the log loss of an $(L + 1)$ -way softmax-based classifier trained to discriminate the (augmented) instance x_i from other images stored in the dictionary queue (represented by their embeddings). Then, θ_q is updated by back-propagation, whereas θ_k is updated with a momentum m : $\theta_k \leftarrow m\theta_k + (1 - m)\theta_q$, where $m \in [0, 1)$. A notable difference of our work from the prevailing meta-learning based approaches is that we randomly sample mini-batches of training images from D_{base} , instead of iteratively constructing episodic training tasks (Finn et al., 2017; Hsu et al., 2018; Khodadadeh et al., 2019; Li et al., 2020; Sung et al., 2018; Snell et al., 2017). This back-to-basic training scheme has proven effective despite being simpler (Tian et al., 2020), especially for an unsupervised D_{base} where category information is missing for episode construction.

After the URL, we freeze θ_q and append a simple classifier f_c to f_q to form a baseline model $F = f_c(f_q)$ for rare disease classification (Fig. 1(b)). Like Tian et al. (2020), we use logistic regression for f_c , whose parameters θ_c are optimized on the support set S .

3.4. Self-distillation of rare disease knowledge

Despite its decent performance, the baseline model completely ignores the precious knowledge about target rare diseases contained in the support set S during representation learning. We hypothesize that a better representation for the classification of the target rare diseases can be learned by fully exploiting this knowledge while at the same time utilizing the big unlabeled data in D_{base} . To do so, we propose to inject target task knowledge extracted from S into the representation learning process via knowledge distillation (Hinton et al., 2015), which can transfer knowledge embedded in a teacher model to a student model. In addition, we adopt the born-again strategy where the teacher and student models have an identical architecture, for its superior performance demonstrated by Furlanello et al. (2018).

The key idea behind our knowledge distillation scheme is that, although D_{base} and D_{rare} comprise disjoint classes, it is common that certain imaging characteristics (e.g., color, shape and/or texture) of the CDNC data are shared by the rare diseases. Therefore, it is feasible to learn rare-disease-distinctive representations and classifiers by *training the networks to classify CDNC instances as rare diseases of similar characteristics*. Mathematically, we use the baseline model F as the teacher model to predict the probabilities of each image x in D_{base} belonging to the rare diseases in D_{rare} (Fig. 1(c)): $p = F(x) = [p_1, \dots, p_N]^T$, where $\sum_{n=1}^N p_n = 1$. Next, we define the pseudo label $y = [y_1, \dots, y_N]^T$ based on p with two alternative strategies: (i) hard labeling where $y_n = 1$ if $n = \arg\max_n p_n$ and 0 otherwise, and (2) soft labeling where $y_n = p_n$. In effect, the first strategy indicates the rare disease that x resembles most, whereas the second reflects the extents of resemblance between x and all the rare diseases in D_{rare} . In addition, we propose a hybrid distilling loss integrating pseudo-label supervised classification and contrastive instance discrimination (Fig. 1(d)). As we will show, the hybrid distillation scheme is important for preventing overfitting to noise and bias in the small support set. Then, adopting the born-again (Furlanello et al., 2018) strategy, we randomly initialize a new query encoder f'_q (parameterized by θ'_q) and a new key encoder f'_k (parameterized by θ'_k) of the same structure as f_q and f_k . In addition, a randomly initialized fully connected layer (followed by softmax), denoted by f'_c and parameterized by θ'_c , is appended to f'_q to compose the student model $F' = f'_c(f'_q)$, which is trained with a hybrid loss \mathcal{L}_{dis} :

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{con}}(x; \theta'_q, \theta'_k) + \mathcal{L}_{\text{cls}}(y, F'(x; \theta'_q, \theta'_c)), \quad (2)$$

where \mathcal{L}_{con} is the contrastive loss defined in Eq. (1), \mathcal{L}_{cls} is the pseudo-label supervised classification loss, and an implicit equal weight is used to balance \mathcal{L}_{con} and \mathcal{L}_{cls} based on our primitive experiments. For hard pseudo labels, the cross-entropy loss is used for \mathcal{L}_{cls} :

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{CE}}(\mathbf{y}, \mathbf{p}') = -\sum_{n=1}^N y_n \log p'_n, \quad (3)$$

where $\mathbf{p}' = F'(x; \theta'_q, \theta'_c) = [p'_1, \dots, p'_N]^T$ is the prediction by the student model. For soft pseudo labels, the Kullback-Leibler divergence loss is used instead:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{KL}}(\mathbf{y} \parallel \mathbf{p}') = \sum_{n=1}^N y_n \log(y_n/p'_n). \quad (4)$$

We will study the impact of choosing hard or soft pseudo labels in Section 4.4.3. Thus, the student model F' (comprising the query encoder f'_q and the fully connected layer f'_c) is updated by back-propagation to minimize the hybrid loss \mathcal{L}_{dis} comprising the contrastive loss \mathcal{L}_{con} and the pseudo label supervised classification loss \mathcal{L}_{cls} , whereas the key encoder f'_k is updated as a momentum version of f'_q .

After distillation, the student model $F' = f'_c(f'_q)$ can be directly used for rare disease classification. One might argue for an alternative way of usage: discarding f'_c but appending a logistic regression classifier fit to the support set to f'_q , just like in the baseline model. However, as confirmed by our comparative experiment (Table 8), direct use of F' performs much better. Lastly, through preliminary experiments, we find that distilling more than once does not bring further improvement. This is consistent with the findings of Tian et al. (2020), where substantial performance improvements were obtained after self-distilling once or twice; then, the performance fluctuated and even decreased with more rounds of distillation. A theoretical analysis of self-distillation by Mobahi et al. (2020) implied that while a few rounds of self-distillation may reduce over-fitting, further rounds may lead to under-fitting and, thus, worse performance. Therefore, we perform the self-distillation only once.

3.5. Adaptive pseudo labels

In practice, the pseudo labels defined by F may not be entirely trustworthy, given the tiny size and potential noise and bias of the support set. This may adversely affect performance of the student model. To alleviate the adverse effect, we further propose adaptive pseudo labels based on the self-adaptive training (Huang et al., 2020) strategy. Concretely, given the prediction \mathbf{p}' by the student model and pseudo labels \mathbf{y} defined above, we combine them as our new training target:

$$\mathbf{y}^{\text{adpt}} = (1 - \alpha) \times \mathbf{y} + \alpha \times \mathbf{p}', \quad (5)$$

where α is a confidence parameter controlling how much we trust the teacher's (student's) knowledge. \mathbf{y}^{adpt} is termed adaptive hard/soft labels depending on \mathbf{y} being hard or soft pseudo labels. Then, we replaced \mathbf{y} in Eq. (2) with the new target \mathbf{y}^{adpt} . Many previous works used a constant α (Huang et al., 2020; Zhang et al., 2019). In the first few epochs, however, the student model lacks reliability—it gradually develops a better ability of rare disease classification as the training goes on. Therefore, we adopt a linear growth rate (Kim et al., 2020b) for α at the i th epoch: $\alpha_i = \alpha_T \times (i/T)$, where α_T is the last-epoch value and set to 0.7 as in Kim et al. (2020b), and T is the total number of epochs. As suggested by the comparative experiments (Table 9), the adaptive hard labels work the best, thus are used in our full model for comparison with other methods.

3.6. Dispersion-aware imbalance correction

Although the above-presented hybrid self-distillation scheme can effectively improve the total accuracy upon the baseline model F , a close examination of the per class accuracy reveals that such improvement often comes from substantial improvements in one or two classes

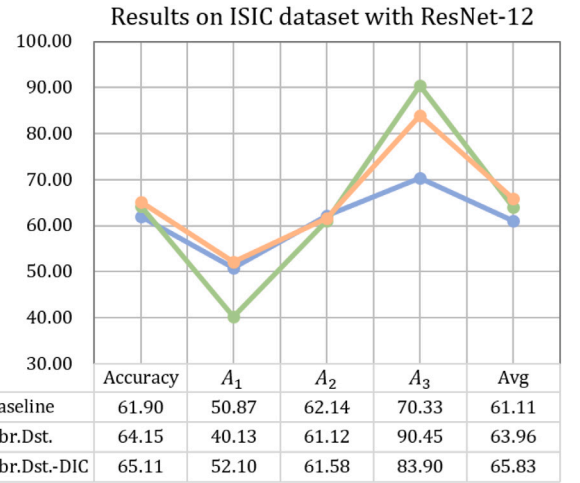


Fig. 2. Example of imbalance between per class accuracies (%) in 3-way 1-shot setting, where A_1 , A_2 , and A_3 are the per class accuracies of three rare diseases, and 'Avg' is their arithmetic mean.

while sacrificing the accuracy of the other(s). For example, Fig. 2 shows that the hybrid distillation (Hbr.Dst.) improves A_3 by over 20% and decreases A_1 by over 10% (A_3 and A_1 are class-wise accuracies of two rare diseases), while improving the total accuracy from 61.90% to 64.15%. Such inter-class imbalance is similar to what is observed in long-tail distributed classification. Wang et al. (2021) proposed a long-tailed classifier that reduced the model bias with a distribution-aware diversity loss (DDL). Specifically, DDL employed a temperature (or concentration) linearly scaled with the class size on each class's logit, to make the model more robust and sensitive to the tail classes. However, in the context of few-shot rare disease classification, the underlying size distribution of the rare diseases can neither be directly estimated from the tiny support set nor reliably from literature-documented incidence rates due to their rareness. A workaround is to estimate the size distribution based on the pseudo labels assigned to the CDNC cases, which, however, reflects the distribution of the CDNC more than that of the rare diseases.

To bypass the difficulty that the class size distribution of the rare diseases cannot be obtained, we instead propose to use the variance of the penultimate layer representation's norm to measure the dispersion of a class. Intuitively, more dispersed classes are more robust to overfitting than less dispersed ones due to their richer diversities. Using this property, we propose a dispersion-aware imbalance correction (DIC) method. Concretely, we apply a variable temperature τ_n^{DIC} to class n 's logit $f'_c(f'_q(x)_n)$ (before softmax):

$$\mathbf{p}^{\text{DIC}} = \text{DIC}(\mathbf{p}') = \text{softmax} \left(\left[\frac{f'_c(f'_q(x)_1)}{\tau_1^{\text{DIC}}}, \dots, \frac{f'_c(f'_q(x)_N)}{\tau_N^{\text{DIC}}} \right] \right), \quad (6)$$

where

$$\tau_n^{\text{DIC}} = \beta_n + 1 - \max_j \beta_j, \quad (7)$$

where

$$\beta_n = \gamma \cdot \frac{v_n}{\sum_{i=1}^N v_i} + (1 - \gamma), \quad (8)$$

$j \in \{1, \dots, N\}$, v_n is the variance of the normalized features $\|f'_q(x)\|_2$ of CDNC samples with pseudo label n , and γ is a reweight factor set to 0.1 in our experiments following Wang et al. (2021). The smaller the variance v_n , the lower the temperature τ_n^{DIC} , and the more sensitive the classification probability \mathbf{p}^{DIC} is to a change in the feature of class n . Thus, the model pays more attention to less dispersed classes and less attention to more dispersed ones. As a result, DIC not only

effectively re-balances the performance on different rare diseases, but also improves the total accuracy (cf. Hbr.Dst.-DIC in Fig. 2).

Combining the adaptive pseudo labels and DIC, the hybrid distillation loss in Eq. (2) now evolves into:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{con}}(x; \theta'_q, \theta'_k) + \mathcal{L}_{\text{cls}}(y^{\text{adpt}}, p^{\text{DIC}}). \quad (9)$$

4. Experiments

4.1. Datasets and evaluation protocol

We conduct experiments on three public datasets: ISIC (Codella et al., 2019; Tschandl et al., 2018),³ Pap-smear (Jantzen et al., 2005), and the optical coherence tomography (OCT) diagnosis of rare retinal diseases dataset (Yoo et al., 2021).

The ISIC 2018 skin lesion classification dataset includes 10,015 dermoscopic images from seven disease categories: melanocytic nevus (6705), benign keratosis (1099), melanoma (1113), basal cell carcinoma (514), actinic keratosis (327), dermatofibroma (115), and vascular lesion (142). The Pap-smear benchmark dataset includes 917 microscopic images of cervical smears across seven classes: superficial squamous epithelial (74), intermediate squamous epithelial (70), columnar epithelial (98), mild squamous non-keratinizing dysplasia (182), moderate squamous non-keratinizing dysplasia (146), severe squamous non-keratinizing dysplasia (197), and squamous cell carcinoma in situ intermediate (150). From both datasets, we simulate tasks of rare disease classification as described below. Following Li et al. (2020) and Singh et al. (2021), we use the four classes with the most cases as the CDNC dataset D_{base} , and the other three as the rare disease dataset D_{rare} .

The OCT diagnosis dataset includes OCT images showing the characteristics of normal retinas (27,110), three major diseases: diabetic macular edema (11,598), drusen (8866), choroidal neovascularization (37,455), and five rare diseases: central serous chorioretinopathy (30), macular telangiectasia (30), macular hole (30), Stargardt disease (16), and retinitis pigmentosa (19). The five rare diseases in the OCT diagnosis dataset are actual clinical rare diseases according to the Orphanet rare disease database (Nguengang Wakap et al., 2020) and a previous review on OCT diagnosis of retinal diseases (Murthy et al., 2016), in contrast to the simulated “rare” diseases in the ISIC and Pap-smear datasets. Naturally, images of the normal retinas and three major diseases compose D_{base} and those of the five rare diseases compose D_{rare} .

To construct a specific rare disease classification task for performance evaluation, K images are randomly sampled for each class in D_{rare} to compose the support set S for an N -way K -shot task ($N = 3$ for the ISIC and Pap-smear datasets, and $N = 5$ for the OCT dataset). Compared to the binary classification tasks (Li et al., 2020), the multi-way evaluation protocol more genuinely reflects the practical clinical use scenario where more than two rare diseases are present, albeit more challenging. As to K , we mainly follow Li et al. (2020) to experiment with 1, 3, and 5 shots in this work. All remaining images in D_{rare} compose the query set Q for performance evaluation. Again, such task construction more genuinely reflects the intended scenario of rare disease classification – where only few examples are available for training a classifier to be applied to all future test cases – than the repeated construction of small Q 's (Li et al., 2020). We sample three random tasks $\mathcal{T} \sim p(\mathcal{T})$ and report the mean and standard deviation of these tasks. Besides total accuracy, we additionally report the accuracy for each rare disease (denoted by A_i) and their average value ($\text{Avg} = \frac{1}{N} \sum_{i=1}^N A_i$) to assess the performance imbalance in the rare disease classification. The Wilcoxon signed-rank test is used for statistical significance testing concerning total and average accuracies.

4.2. Implementation

The PyTorch (Steiner et al., 2019) framework (1.4.0) is used for experiments. We use the ResNet-12 (Lee et al., 2019; Ravichandran et al., 2019; Tian et al., 2020) architecture as backbone network for f_q and f_k . We train the networks for 200 epochs with a mini-batch size of 16 images on 4 T V100 GPUs. We adopt the stochastic gradient descent optimizer with a momentum of 0.9 and a weight decay of 0.0001. The learning rate is initialized to 0.03 and decays at 120th and 160th epochs by a factor of 0.1. The feature dimension of the encoded representation is 128, and the number of negatives in the memory bank (He et al., 2020) is 1280. The temperature τ in Eq. (1) is set to 0.07 and momentum m is set to 0.999 as in He et al. (2020). All images are resized to 224×224 pixels. Online data augmentation including random cropping, flipping, color jittering, and blurring (Chen et al., 2020a) is performed. The source code is available at: <https://github.com/jinghanSunn/Hybrid-Representation-Learning-Approach-for-Rare-Disease-Classification>.

4.3. Comparison to SOTA methods

According to the labeling status of the base dataset and genre of the methodologies, all the compared methods are grouped into four quadrants: (i) supervised meta-learning (SML) including MAML (Finn et al., 2017), Relation Networks (Sung et al., 2018), Prototypical Networks (Snell et al., 2017), and DAML (Li et al., 2020), (ii) unsupervised meta-learning (UML) including UMTRA (Khodadadeh et al., 2019) and CACTUs (CCTs) (Hsu et al., 2018), (iii) supervised representation learning (SRL; with or without self-distillation) (Tian et al., 2020), and (iv) URL including SimCLR (Chen et al., 2020a), MoCo_v2 (Chen et al., 2020b), MoCo_v1 (He et al., 2020) (composing the baseline model in this work), our proposed hybrid distillation (Hbr.Dst.) without imbalance correction, and with the distribution-aware diversity loss (DDL) (Wang et al., 2021) and our dispersion-aware imbalance correction (DIC), respectively. Note that as our CDNC dataset is unlabeled, we use the predicted hard pseudo labels to generate the temperature parameter for DDL. These methods cover a wide range of the latest advances in FSL for image classification. For reference purposes, we also show the results of training a classifier from scratch solely on the support set S . Besides the ResNet-12 backbone, we additionally show the results using the 4-conv-block backbone (Vinyals et al., 2016), considering its prevalent usage in the FSL literature (Finn et al., 2017; Sung et al., 2018; Snell et al., 2017; Khodadadeh et al., 2019; Hsu et al., 2018). For all compared methods, we optimize their performance via empirical parameter tuning.

4.3.1. Results on the ISIC dataset

The results on the ISIC dataset are shown in Table 1 (ResNet-12) and Table 2 (the 4-conv-block backbone), on which we make the following observations. First, the representation learning based methods generally achieve better performance than the meta-learning based irrespective of the labeling status of D_{base} , which is consistent with the findings in the natural image domain (Tian et al., 2020). Second, the URL-based methods surprisingly outperform the SRL-based in most circumstances. Especially, the baseline model presented in this work (URL with MoCo_v1 He et al. (2020)) generally outperforms the SRL plus self-distillation (SRL-distill) (Tian et al., 2020), often by large margins. Third, our proposed hybrid distillation (Hbr.Dst.) approach brings further improvements upon the baseline model in both total and average accuracies (e.g., $\sim 1\%$ – 2% and $\sim 2\%$ – 3% with ResNet-12 in Table 1, respectively). However, we also note that the imbalance between A_1 – A_3 is more severe for our Hbr.Dst. than for SRL-distill, indicating the need for correction. Lastly, incorporating the proposed DIC into our Hbr.Dst. apparently alleviates the imbalance in all settings, resulting in obvious improvement in average accuracy while maintaining comparable total accuracy. Eventually, our ResNet-12 model

³ <https://challenge2018.isic-archive.com/task3/>.

Table 1

Evaluation results (in %) and comparison with SOTA FSL methods on the ISIC dataset (ResNet-12 backbone). Standard deviation of the accuracy is parenthesized. A_1 , A_2 , and A_3 correspond to dermatofibroma, actinic keratosis, and vascular lesion, respectively.

Method	$(N, K) = (3, 1)$					$(N, K) = (3, 3)$					$(N, K) = (3, 5)$				
	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg
Train scratch	37.74* (1.07)	35.17	41.31	30.10	35.19*	39.76* (0.88)	37.57	42.91	34.21	38.23*	45.36* (3.76)	44.34	48.78	38.17	43.76*
<i>Supervised meta-learning (SML)</i>															
MAML	47.49* (5.38)	41.09	48.01	49.30	46.13*	55.55* (3.12)	48.41	59.24	52.69	53.45*	58.94* (2.59)	56.31	59.70	52.99	56.33*
RelationNet	46.10* (4.80)	38.19	49.45	42.76	44.13*	47.29* (2.77)	46.62	49.49	42.72	46.28*	55.71* (3.30)	50.77	57.59	55.25	54.54*
ProtoNet	35.18* (3.12)	30.18	38.03	32.68	33.63*	38.59* (1.91)	32.61	41.88	35.77	36.75*	42.45* (2.45)	38.74	46.45	34.40	39.86*
DAML	50.05* (5.18)	47.28	54.91	41.07	47.75*	55.57* (3.55)	48.25	57.09	56.59	53.98*	59.44* (3.17)	52.47	61.83	59.43	57.91*
<i>Unsupervised meta-learning (UML)</i>															
UMTRA	45.88* (3.63)	40.49	46.99	47.69	45.06*	41.44* (4.37)	45.69	52.77	52.36	50.27*	57.33* (1.76)	52.30	58.22	59.13	56.69*
CCTs-MAML	42.98* (2.91)	34.44	46.19	42.49	41.04*	44.44* (3.35)	39.51	45.97	44.88	43.45*	48.11* (4.20)	43.23	49.61	47.15	46.95*
CCTs-ProtoNets	42.67* (2.43)	46.43	38.17	45.60	43.40*	45.00* (3.26)	45.23	47.11	39.93	44.09*	47.95* (3.52)	44.15	51.41	47.10	47.55*
<i>Supervised representation learning (SRL)</i>															
SRL-simple	54.45* (5.82)	51.49	55.24	55.05	53.93*	61.31* (6.31)	54.09	63.58	61.84	59.84*	70.53* (2.17)	61.50	74.41	68.70	68.20*
SRL-distill	55.43* (7.36)	50.74	55.16	59.87	55.26*	64.92* (6.00)	58.42	68.25	62.43	63.03*	72.78* (1.67)	62.47	78.11	68.56	69.71*
<i>Unsupervised representation learning (URL)</i>															
SimCLR	52.43* (5.01)	34.94	54.35	62.17	50.49	63.82* (3.70)	46.77	75.53	50.28	57.53*	70.18* (1.76)	60.76	75.28	65.77	67.27*
MoCo_v2	59.95* (4.73)	41.97	63.23	66.94	57.38*	70.84* (2.91)	52.71	70.76	85.67	69.71*	75.80* (1.85)	58.76	80.36	78.80	72.64*
MoCo_v1	61.90* (2.92)	50.87	62.14	70.33	61.11*	74.92 (2.96)	55.27	75.73	88.87	73.29*	79.01* (2.00)	57.00	84.12	84.67	75.26*
Hbr.Dst.	64.15 (2.86)	40.31	61.12	90.45	63.96	75.82 (2.47)	60.27	74.59	91.24	75.37	81.16 (2.60)	69.09	85.31	81.09	78.50
Hbr.Dst.-DDL	64.65 (2.91)	47.39	61.50	85.91	64.93	75.29 (2.88)	58.51	76.19	86.72	73.81	82.02 (2.95)	62.40	88.31	83.01	77.91
Hbr.Dst.-DIC	65.11 (2.90)	52.01	61.58	83.90	65.83	76.48 (3.11)	64.31	77.05	84.96	75.44	82.02 (2.72)	70.15	85.78	82.74	79.56

* $p < 0.05$ for comparison with the proposed Hbr.Dst.-DIC.

Table 2

Evaluation results (in %) and comparison with SOTA FSL methods on the ISIC dataset (the 4-conv-block backbone). Standard deviation of the accuracy is parenthesized. A_1 , A_2 , and A_3 correspond to dermatofibroma, actinic keratosis, and vascular lesion, respectively.

Method	$(N, K) = (3, 1)$					$(N, K) = (3, 3)$					$(N, K) = (3, 5)$				
	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg
Train scratch	35.91* (4.11)	28.94	38.95	34.53	34.41*	38.95* (4.03)	39.84	39.91	32.89	37.55*	43.13* (5.50)	36.87	47.09	38.36	45.85*
<i>Supervised Meta-learning (SML)</i>															
MAML	46.20* (4.22)	38.66	48.46	46.82	44.65*	51.56* (3.03)	44.94	53.51	52.53	50.33*	55.73* (3.62)	48.54	56.72	56.96	54.07*
RelationNet	36.11* (4.72)	30.07	38.66	35.87	34.87*	42.63* (4.57)	43.51	44.65	35.5	41.22*	49.98* (5.81)	40.16	49.19	50.70	46.68*
ProtoNet	34.97* (1.69)	26.85	37.65	35.34	33.28*	36.38* (1.91)	42.25	30.87	39.05	37.39*	37.95* (2.98)	33.85	41.90	31.84	35.86*
DAML	46.70* (2.74)	46.76	48.21	43.40	46.12*	53.91* (3.23)	42.35	56.86	54.26	51.16*	56.11* (2.39)	52.63	58.84	51.90	54.46*
<i>Unsupervised Meta-learning (UML)</i>															
UMTRA	41.45* (2.47)	34.55	38.72	45.44	39.57*	45.76* (2.78)	36.25	45.65	47.21	43.04*	54.16* (2.20)	50.50	57.62	46.02	51.38*
CCTs-MAML	39.45* (3.42)	32.42	41.91	39.47	37.93*	42.20* (2.47)	36.63	42.09	45.56	41.43*	45.31* (1.69)	35.88	48.88	45.29	43.35*
CCTs-ProtoNets	33.24* (2.12)	28.89	34.74	33.21	32.28*	35.08* (1.70)	30.35	38.97	30.46	33.26*	36.86* (1.95)	26.85	39.00	34.07	33.31*
<i>Supervised Representation Learning (SRL)</i>															
SRL-simple	45.08* (6.59)	38.59	46.06	47.9	44.18*	57.50* (4.48)	52.77	64.52	41.86	49.05*	60.62* (4.34)	51.55	67.99	50.91	56.82*
SRL-distill	50.14* (3.74)	43.85	54.90	43.84	47.53*	58.20* (4.15)	55.28	57.05	64.53	58.95*	62.07* (5.11)	50.00	69.72	55.25	58.32*
<i>Unsupervised Representation Learning (URL)</i>															
SimCLR	48.07* (6.46)	24.05	54.27	53.19	43.84*	60.78* (4.18)	20.17	75.57	60.99	52.24*	66.25* (5.36)	41.07	76.31	63.39	60.26*
MoCo_v2	48.88* (8.31)	13.66	71.91	60.37	50.78*	59.71* (2.54)	21.92	65.44	75.17	54.18*	62.91* (2.04)	35.25	77.72	50.00	54.32*
MoCo_v1	53.36 (8.89)	28.94	60.11	57.53	48.86*	67.48 (6.06)	18.42	78.52	79.5	58.81*	69.94 (5.67)	32.78	84.13	66.02	60.98*
Hbr.Dst.	54.72 (8.46)	35.32	60.12	57.98	51.14	69.14 (5.36)	39.17	83.10	60.99	61.09	70.26 (4.97)	38.39	82.72	67.35	62.82
Hbr.Dst.-DDL	54.89 (7.01)	49.78	57.95	51.95	53.23	67.24 (6.19)	50.71	75.27	61.87	62.62	69.08 (4.92)	41.81	82.41	60.29	61.50
Hbr.Dst.-DIC	55.30 (7.72)	55.10	56.27	53.23	54.87	69.33 (4.69)	55.71	77.62	61.01	64.78	70.35 (5.73)	52.72	80.21	61.83	64.92

* $p < 0.05$ for comparison with the proposed Hbr.Dst.-DIC.

achieves a remarkable accuracy of 82.02% in the 5-shot setting without any label of the base dataset. On the other hand, incorporating the DDL (Wang et al., 2021) does not effectively address the performance imbalance issue and often reduces the average accuracy.

To further investigate whether the improvements sustain more support data, we conduct extra experiments with $K = 10$ and 20 using the ResNet-12 backbone (Table 4). The results confirm that our proposed hybrid approach still wins over the baseline model in both metrics and settings, and the incorporation of DIC brings further improvements again on both metrics and settings, although the absolute differences become less prominent compared to those with less support data, as expected.

4.3.2. Results on the Pap-smear dataset

The results on the Pap-smear dataset are shown in Table 3 (ResNet-12) and Table 5 (the 4-conv-block backbone). The results present similar trends to those on the ISIC dataset with the proposed Hbr.Dst.-DIC achieving the highest total and average accuracies in all K settings and using both backbones, again demonstrating (1) the representation power of the URL, (2) the efficacy of the hybrid distillation in injecting knowledge from the limited support data, and (3) the effective correction of performance imbalance by DIC. Notably, our Hbr.Dst.-DIC (ResNet-12 variant) yields an accuracy of 84.05% in the 5-shot setting without any label of the base dataset.

Table 3

Evaluation results (in %) and comparison with SOTA FSL methods on the Pap-smear dataset (ResNet-12 backbone). Standard deviation of the accuracy is parenthesized. A_1 , A_2 , and A_3 correspond to columnar epithelial, superficial squamous epithelial, and intermediate squamous epithelial, respectively.

Method	$(N, K) = (3, 1)$					$(N, K) = (3, 3)$					$(N, K) = (3, 5)$				
	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg
Train scratch	28.14* (1.77)	27.28	27.87	29.66	28.27*	39.51* (1.19)	37.81	39.40	42.06	39.76*	47.56* (1.28)	46.42	46.42	47.52	47.86*
<i>Supervised Meta-learning (SML)</i>															
MAML	32.14* (1.41)	31.34	32.05	33.39	32.26*	56.53* (1.14)	54.87	56.73	58.70	56.77*	59.24* (1.00)	62.90	51.90	65.22	61.25*
RelationNet	30.49* (1.01)	29.28	30.48	32.22	30.66*	55.66* (1.75)	53.46	56.43	57.99	55.96*	60.71* (1.10)	62.08	52.08	64.03	59.69*
ProtoNet	32.12* (2.01)	30.69	30.87	32.00	31.19*	43.36* (1.06)	44.52	42.99	42.11	43.21	49.42* (1.00)	48.34	48.34	49.11	49.14*
DAML	33.09* (1.48)	30.50	33.71	36.09	33.43*	57.45* (1.37)	52.42	60.15	61.77	58.11*	63.30* (1.84)	60.30	60.30	67.05	59.62*
<i>Unsupervised Meta-learning (UML)</i>															
UMTRA	32.60* (2.01)	32.79	31.08	33.97	32.61*	56.17* (2.98)	55.20	56.38	57.31	56.30*	63.28* (1.76)	62.13	64.26	60.01	62.13*
CCTs-MAML	31.44* (2.00)	32.59	30.05	31.29	31.31*	55.29* (2.04)	56.09	55.28	54.17	55.18*	63.64* (1.07)	62.91	60.82	63.77	62.50*
CCTs-ProtoNets	31.02* (1.09)	30.84	32.22	29.98	31.01*	52.98* (2.88)	51.05	52.58	56.18	53.27*	62.88* (2.09)	60.68	61.10	64.19	61.99*
<i>Supervised Representation Learning (SRL)</i>															
SRL-simple	52.96* (2.11)	50.52	54.16	55.15	53.28*	69.02* (1.91)	66.02	68.13	74.24	69.46*	75.38* (2.87)	71.74	75.73	75.78	74.42*
SRL-distill	57.32* (1.79)	57.41	54.65	60.00	57.35*	68.84* (1.88)	62.59	75.94	70.23	69.59*	76.41* (1.81)	76.17	74.09	74.46	74.46*
<i>Unsupervised Representation Learning (URL)</i>															
MoCo_v1	62.58* (3.51)	59.72	59.54	69.85	63.04*	71.05* (2.37)	59.41	74.94	83.52	72.62*	81.90* (2.37)	75.20	83.88	84.75	83.39*
Hbr.Dst.	64.97 (2.30)	56.64	70.77	70.61	66.01	72.32 (2.92)	62.72	79.02	78.91	73.55	83.08 (2.84)	71.38	88.81	89.36	83.18
Hbr.Dst.-DDL	64.38 (2.55)	55.74	72.01	68.52	65.42	73.28 (2.88)	62.52	80.01	81.50	74.68	82.62 (3.09)	73.41	86.59	87.04	82.35
Hbr.Dst.-DIC	66.18 (2.15)	59.37	70.66	72.11	67.05	74.16 (2.43)	64.89	81.87	79.21	75.32	84.05 (2.09)	76.38	87.88	86.25	83.50

* $p < 0.05$ for comparison with the proposed Hbr.Dst.-DIC.

Table 4

Evaluation results with more support data (i.e., larger K values) on the ISIC dataset (ResNet-12 backbone). Format: mean (standard deviation).

	$(N, K) = (3, 10)$		$(N, K) = (3, 20)$	
	Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)
MoCo_v1	82.66 (3.95)	81.73 (3.02)	86.33 (4.53)	84.19 (4.04)
Hbr.Dst.	83.65 (3.41)	82.09 (2.89)	86.91 (2.37)	84.33 (3.77)
Hbr.Dst.-DIC	83.80 (3.44)	82.77 (3.17)	87.25 (3.12)	85.00 (3.42)

4.3.3. Results on the OCT dataset

The results on the OCT dataset are shown in Table 6. Unlike the ISIC and Pap-smear datasets, this dataset includes real clinical rare diseases instead of simulated ones. It is also more challenging considering the higher number of rare diseases to differentiate and the much fewer testing samples than the ISIC dataset. Consequently, many methods experience significant performance drops compared to Tables 1 and 3, especially the meta-learning ones. In contrast, the representation learning based methods are more robust, yielding relatively comparable performance on the OCT dataset. Notably, our full model Hbr.Dst.-DIC again achieves the highest total and average accuracies in all K settings. These results suggest that our assumption that CDNCs and rare diseases share common characteristics that can be learned on the former and effective on the latter is valid, and that our proposed method is also effective, for real clinical rare diseases.

4.4. Ablation study

4.4.1. Efficacy of building components

Using the ResNet-12 backbone on the ISIC dataset, we probe the effect of the proposed hybrid distillation by comparing the performance of no distillation, distilling with only \mathcal{L}_{cls} or both \mathcal{L}_{cls} and \mathcal{L}_{con} . In addition, we experiment with a variant of \mathcal{L}_{dis} where \mathcal{L}_{cls} is replaced by an L1 loss \mathcal{L}_{reg} to directly regress the features output by f_q . Table 7 shows that distilling with \mathcal{L}_{cls} alone (row (b)) brings slight improvement upon the baseline model with no distillation (row (a)), suggesting the efficacy of injecting rare disease knowledge into representation learning. Yet, distilling with the proposed hybrid loss achieves further substantial improvement (row (c)), demonstrating the mutual benefit of combining \mathcal{L}_{cls} and \mathcal{L}_{con} . On the other hand, substituting \mathcal{L}_{reg} for \mathcal{L}_{cls} (row (d)) gives performance similar to the baseline model, implying that it is the

distilled knowledge about the rare diseases that matters, rather than the distillation procedure. Lastly, the addition of DIC (the full model) leads to the optimal performance in this work. These results confirm the efficacy of the proposed hybrid distillation and DIC.

4.4.2. Alternative strategies of applying distilled student model

In Table 8, we compare alternative strategies of applying the distilled student model: directly using the student model $F' = f'_c(f'_q)$ ("Direct"), or replacing f'_c with a logistic regression classifier fit to the support set ("LR"). As we can see, the performance of the Direct strategy is apparently better than that of the LR for both evaluated metrics and all few-shot settings. This is because, the knowledge about the rare diseases is distilled into not only f'_q but also f'_c , thus discarding the latter results in performance degradation.

4.4.3. Choices of pseudo labels

Table 9 shows how different choices of the pseudo labels affect the performance. As we can see, hard labels are better than soft labels in all settings. In addition, the adaptive pseudo labels work better than the vanilla hard pseudo labels, increasing the total and average accuracies by up to 2.64% and 2.47%, respectively. Therefore, it is most effective to use the adaptive hard labels for supervision in Eq. (9).

4.5. In-depth analyses

4.5.1. Model training strategy

In our problem setting, there are only several samples for each rare disease, and all of them are used as the support set for maximal utilization in model training. As a result, there is no validation dataset for model selection during training. Hence, it is critical to ensure the baseline unsupervised representation learning model (Fig. 1(b)) and the hybrid distillation model (Fig. 1(d)) are trained well. In this work, we adopt a common strategy in the contrastive learning literature (Chen et al., 2020b, 2021a), where the training loss is monitored to determine the model convergence during training, and use the last-epoch model for performance evaluation and comparison. Fig. 3 plots the training curve of \mathcal{L}_{con} for the baseline model (MoCo_v1). For an in-depth analysis and only for the purpose of analysis, the evolution of the testing-(query-) set accuracies under different K values during training is also drawn. As we can see, the loss appears to have well converged after 175 epochs, and we train for a total of 200 epochs. The accuracy also

Table 5

Evaluation results (in %) and comparison with SOTA FSL methods on the Pap-smear dataset (the 4-conv-block backbone). Standard deviation of the accuracy is parenthesized. A_1 , A_2 , and A_3 correspond to columnar epithelial, superficial squamous epithelial, and intermediate squamous epithelial, respectively.

Method	$(N, K) = (3, 1)$					$(N, K) = (3, 3)$					$(N, K) = (3, 5)$				
	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg
Train scratch	30.10* (3.43)	30.88	31.94	27.08	29.97*	34.32* (3.39)	30.00	38.29	39.39	35.89*	38.33* (3.60)	30.55	37.50	48.52	38.86*
<i>Supervised Meta-learning (SML)</i>															
MAML	42.06* (3.53)	48.48	37.14	42.55	42.72*	47.39* (4.46)	51.38	47.91	41.17	46.82*	51.13* (4.24)	48.61	50.00	52.94	50.52*
RelationNet	35.47* (1.54)	32.35	41.66	33.33	35.78*	46.94* (1.47)	43.05	47.91	51.47	47.48*	48.86* (3.47)	43.05	53.12	50.00	48.72*
ProtoNet	37.04* (3.74)	33.82	45.83	32.29	37.31*	40.88* (1.08)	47.05	40.21	32.81	40.02*	46.89* (1.59)	50.00	42.55	43.93	45.49*
DAML	42.77* (3.50)	42.64	36.11	50.00	42.92*	49.86* (4.40)	52.85	45.74	50.00	49.53*	51.31* (3.67)	41.66	52.08	61.76	51.83*
<i>Unsupervised Meta-learning (UML)</i>															
UMTRA	40.23* (2.61)	25.00	43.05	58.75	42.27*	51.44* (2.47)	37.50	77.08	44.11	52.90*	61.20* (2.08)	54.16	69.79	58.82	60.92*
CCTs-MAML	39.43* (3.95)	27.94	40.83	54.16	40.98*	51.88* (1.14)	60.00	12.76	81.81	51.52*	58.05* (3.68)	67.14	26.59	74.24	55.99*
CCTs-ProtoNets	42.09* (2.37)	50.00	31.94	41.66	41.20*	46.39* (2.48)	45.83	47.91	45.58	46.44*	51.62* (2.96)	54.16	52.08	44.11	50.12*
<i>Supervised Representation Learning (SRL)</i>															
SRL-simple	46.76* (7.09)	25.00	52.77	53.95	47.97*	57.63* (5.78)	44.28	62.76	71.21	59.42*	65.63* (3.07)	58.82	68.47	68.75	65.35*
SRL-distill	46.39* (6.26)	30.88	54.16	60.08	48.37*	60.36* (4.50)	48.57	65.95	71.21	61.91*	67.45* (2.74)	57.35	70.65	75.00	67.67*
<i>Unsupervised Representation Learning (URL)</i>															
MoCo_v1	54.09 (7.12)	47.35	54.16	63.54	55.02*	63.32 (7.23)	60.42	66.95	63.63	63.67*	73.09 (6.55)	66.17	78.26	73.43	72.62*
Hbr.Dst.	55.16 (6.37)	46.82	51.11	71.25	56.39	64.02 (8.66)	50.00	71.27	76.33	65.87	72.93 (4.30)	72.76	72.82	68.75	71.44
Hbr.Dst.-DDL	56.61 (6.19)	51.17	55.88	65.08	57.38	65.16 (7.02)	52.00	72.34	76.30	66.88	73.52 (4.89)	64.70	78.26	77.12	73.36
Hbr.Dst.-DIC	58.12 (6.78)	55.05	55.55	65.20	58.60	66.37 (7.33)	62.12	66.02	72.81	66.98	75.86 (4.84)	76.41	77.34	77.75	75.17

* $p < 0.05$ for comparison with the proposed Hbr.Dst.-DIC.

Table 6

Evaluation results (in %) and comparison with SOTA FSL methods on the OCT dataset (ResNet-12 backbone). Standard deviation of the accuracy is parenthesized. A_1 – A_5 correspond to central serous chorioretinopathy, macular telangiectasia, macular hole, Stargardt disease, and retinitis pigmentosa, respectively.

Method	$(N, K) = (5, 1)$							$(N, K) = (5, 3)$							$(N, K) = (5, 5)$						
	Accuracy	A_1	A_2	A_3	A_4	A_5	Avg	Accuracy	A_1	A_2	A_3	A_4	A_5	Avg	Accuracy	A_1	A_2	A_3	A_4	A_5	Avg
Train scratch	21.48* (0.98)	21.48	10.04	26.35	29.11	20.94	21.58*	24.79* (0.75)	18.89	24.79	18.73	23.72	26.92	22.61*	23.14* (0.98)	19.26	23.33	23.14	12.30	31.72	21.95*
<i>Supervised Meta-learning (SML)</i>																					
MAML	45.45* (2.01)	45.45	45.20	53.17	51.28	37.26	46.48*	48.54* (1.51)	39.61	48.54	75.48	49.38	39.66	50.53*	49.88* (0.92)	16.67	44.23	49.88	82.07	53.18	49.21*
RelationNet	31.40* (1.55)	31.40	37.04	16.67	34.62	23.08	28.56*	34.71* (1.50)	32.28	34.71	38.71	35.71	33.33	34.95*	37.19* (1.77)	9.09	28.70	37.19	24.14	42.31	28.29*
ProtoNet	33.33* (1.08)	33.33	38.03	32.45	34.87	23.52	32.44*	35.81* (0.98)	27.94	35.81	23.87	40.55	38.92	33.42*	36.36* (1.54)	14.92	31.88	36.36	27.90	30.25	28.26*
DAML	41.59* (2.59)	41.59	35.30	51.47	39.60	40.79	41.75*	47.11* (1.98)	35.37	47.11	57.67	48.24	45.98	46.87*	49.60* (2.04)	4.76	38.98	49.60	83.26	54.32	46.18*
<i>Unsupervised Meta-learning (UML)</i>																					
UMTRA	38.84* (2.05)	38.84	23.33	37.04	41.38	55.17	39.15*	46.28* (1.86)	33.29	46.28	54.84	46.43	50	46.17*	52.06* (1.79)	0.00	37.59	52.06	54.84	50.00	38.90*
CCTs-MAML	38.84* (1.97)	38.84	33.85	35.53	36.51	50.87	39.12*	41.04* (1.55)	32.69	41.04	31.43	38.54	42.68	37.28*	43.52* (1.59)	6.20	34.12	43.52	32.69	49.29	33.16*
CCTs-ProtoNets	36.36* (2.45)	36.36	29.00	35.35	37.46	42.86	36.20*	38.56* (2.75)	31.90	38.56	24.16	46.12	37.59	35.67*	41.32* (2.50)	9.38	32.80	41.32	32.26	46.43	32.44*
<i>Supervised Representation Learning (SRL)</i>																					
SRL-simple	53.83* (3.95)	53.83	65.52	53.66	46.11	54.84	54.79*	57.50* (1.99)	46.43	57.50	95.40	53.85	27.38	56.11*	66.66* (1.56)	19.04	53.90	66.66	96.30	63.89	59.96*
SRL-distill	56.55* (3.95)	56.55	74.12	56.23	49.84	54.44	58.24*	59.03* (3.30)	48.42	59.03	94.25	56.41	44.05	60.43*	69.97* (2.22)	21.42	55.37	69.97	96.30	66.67	61.94*
<i>Unsupervised Representation Learning (URL)</i>																					
MoCo_v1	64.87* (3.90)	64.87	96.47	56.73	64.38	65.66	69.62	69.88* (2.15)	96.38	61.00	68.68	68.50	44.27	67.77*	75.19* (2.41)	97.53	68.05	75.64	71.79	56.12	73.83*
Hbr.Dst.	66.85 (2.04)	68.32	67.74	70.87	77.78	66.97	70.34	71.89 (3.63)	67.05	75.85	75.64	63	80.69	72.45	76.84 (2.81)	74.07	75	73.07	80.77	72.90	75.16
Hbr.Dst.-DDL	67.03 (2.02)	65.04	66.19	69.55	75.55	65.86	68.44	72.91 (1.99)	75.76	73.92	73.25	70.49	70.45	72.77	76.42 (2.81)	80.65	76.39	79.74	65.38	81.05	76.64
Hbr.Dst.-DIC	69.59 (2.05)	69.46	65.12	73.19	70.10	73.15	70.49	73.58 (1.38)	75.75	79.93	74.59	63.16	75.30	73.75	77.96 (1.49)	75.71	79.16	78.46	78.20	78.62	78.03

* $p < 0.05$ for comparison with the proposed Hbr.Dst.-DIC.

Table 7

Ablation study on different components of the proposed method (with ResNet-12 backbone and adaptive hard labels on the ISIC dataset). Format: mean (standard deviation).

Ablation	L_{dis}			DIC	$(N, K) = (3, 1)$		$(N, K) = (3, 3)$		$(N, K) = (3, 5)$	
	L_{con}	L_{cls}	L_{reg}		Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)
(a)	×	×	×	×	61.90 (2.92)	61.11 (2.78)	74.92 (2.96)	73.29 (3.17)	79.01 (2.00)	75.26 (2.76)
(b)	×	✓	×	×	63.70 (3.39)	61.09 (4.39)	74.92 (2.10)	73.05 (3.20)	80.24 (1.61)	75.53 (2.57)
(c)	✓	✓	×	×	64.15 (2.86)	63.96 (2.53)	75.82 (2.70)	75.37 (2.52)	81.16 (2.60)	78.50 (2.10)
(d)	✓	×	✓	×	62.20 (5.18)	62.02 (4.78)	74.43 (2.80)	73.19 (3.32)	79.14 (2.09)	74.42 (2.16)
Full	✓	✓	×	✓	65.11 (2.90)	65.83 (3.87)	76.48 (3.11)	75.44 (3.02)	82.02 (2.72)	79.56 (3.05)

Table 8

Performance comparison of alternative strategies of applying the distilled student model (with the ResNet-12 backbone and adaptive hard labels) on the ISIC dataset. “Direct” means directly using the student model $F' = f'_c(f'_q)$, and “LR” means replacing f'_c with a logistic regression classifier fit to the support set. Format: mean (standard deviation).

Method	Classifier	$(N, K) = (3, 1)$		$(N, K) = (3, 3)$		$(N, K) = (3, 5)$	
		Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)
Hbr.Dst. (ours)	LR	62.86 (2.70)	61.17 (3.14)	74.45 (2.59)	73.62 (2.88)	80.55 (1.94)	76.39 (2.96)
	Direct	64.15 (2.86)	63.96 (2.53)	75.82 (2.47)	75.37 (2.52)	81.16 (2.60)	78.50 (2.10)
Hbr.Dst.-DIC (ours)	LR	62.47 (2.53)	62.87 (3.18)	74.95 (2.09)	74.95 (1.87)	81.01 (1.57)	77.31 (3.59)
	Direct	65.11 (2.90)	65.83 (3.87)	76.48 (3.11)	75.44 (3.02)	82.02 (2.72)	79.56 (3.05)

Table 9

Performance comparison with different choices of the pseudo labels on the ISIC dataset (based on Hbr.Dst-DIC with the ResNet-12 backbone).
Format: mean (standard deviation).

Method	$(N, K) = (3, 1)$		$(N, K) = (3, 3)$		$(N, K) = (3, 5)$	
	Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)	Accuracy (%)	Avg (%)
Soft	61.79 (4.31)	64.83 (3.76)	74.26 (3.07)	72.67 (3.55)	80.84 (2.58)	77.07 (2.86)
Hard	62.47 (2.56)	65.05 (3.26)	75.02 (3.95)	72.97 (3.73)	81.37 (2.69)	78.40 (2.55)
Adaptive soft	63.68 (3.71)	65.40 (3.98)	75.05 (3.33)	74.58 (3.13)	81.89 (2.53)	78.95 (2.03)
Adaptive hard	65.11 (2.90)	65.83 (3.87)	76.48 (3.11)	75.44 (3.02)	82.02 (2.72)	79.56 (3.05)

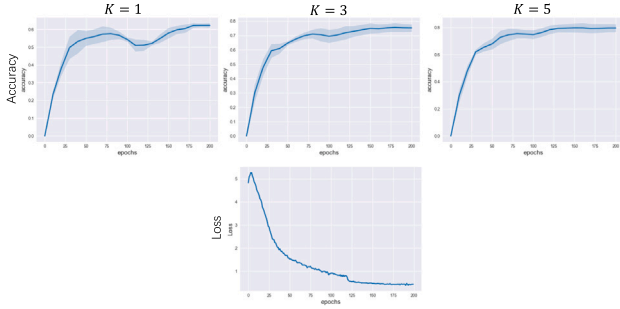


Fig. 3. Testing accuracy and training loss (\mathcal{L}_{con} in Eq. (1)) curves of the baseline model (MoCo_v1) on the ISIC dataset. The testing accuracy curves are only visualized for in-depth analyses, not model selection. For each K value, the accuracy plot shows the mean total accuracy (the dark blue central line) of three runs with the corresponding span overlaid (the light blue shaded strip). Because the same pretrained backbone network is used for different K values and repeated runs, there is only one loss curve with no overlaid span.

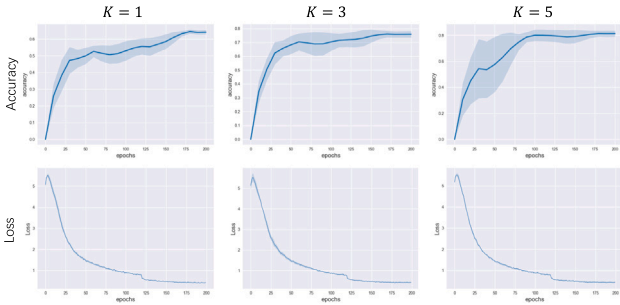


Fig. 4. Training loss (\mathcal{L}_{dis} in Eq. (9)) and testing accuracy curves of our full model (Hbr.Dst-DIC) on the ISIC dataset. The testing accuracy curves are only visualized for in-depth analyses, not model selection. For each K value, the accuracy/loss plot shows the mean values (the dark blue central line) of three runs with the corresponding span overlaid (the light blue shaded strip); note the loss spans are narrow.

converges with the loss, and no apparent evidence of overfitting is observed. These observations validate our loss-based training strategy.

Similarly, Fig. 4 plots the joint training loss \mathcal{L}_{dis} and testing accuracy curves of the proposed hybrid distillation model (Hbr.Dst-DIC). It is apparent that both the training loss and testing accuracy have converged synchronously after ~ 190 epochs, and no evidence of overfitting shows, validating the strategy to judge the convergence of the hybrid distillation model based on the joint training loss.

4.5.2. Quality of pseudo labels and representations

We examine the consistency of the predicted pseudo labels assigned to the CDNC instances (cf. Fig. 1(c)) by looking at their distributions across different shots and runs. Fig. 5 shows the distributions for the ISIC dataset. As we can see, the relative distributions of the three pseudo labels are consistent across both shots and runs, with actinic keratosis (44%–57%) > vascular lesion (30%–40%) > dermatofibroma (11%–22%). The percentage spans for all three pseudo labels are small, within the range of 10%–13%. These indicate that our baseline

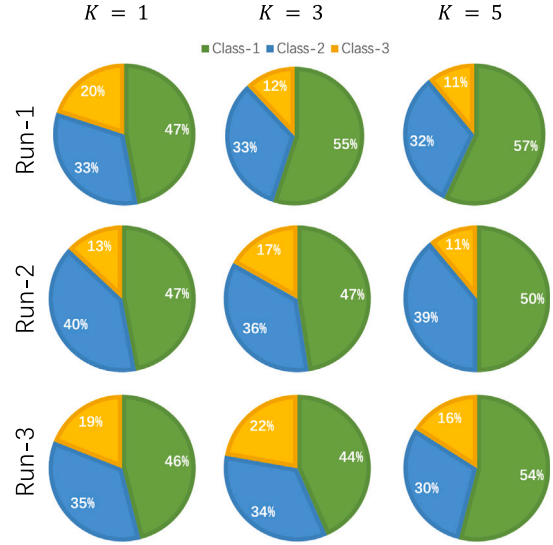


Fig. 5. Class distributions of the predicted pseudo labels for the CDNC instances of the ISIC dataset across different shots and runs with the ResNet-12 backbone. Class-1, Class-2, and Class-3 correspond to actinic keratosis, vascular lesion, and dermatofibroma, respectively.

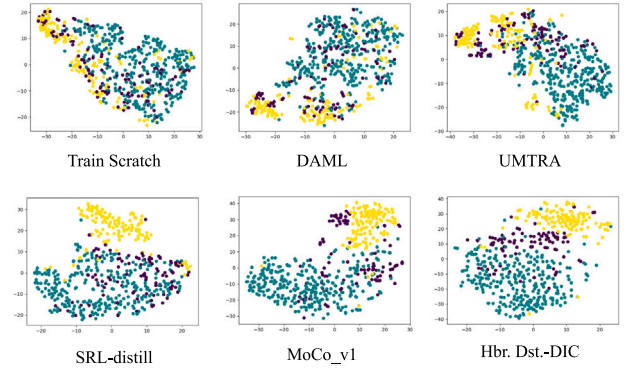


Fig. 6. Visualization of 5-shot representations learned by various methods via t-SNE display of rare disease features on the ISIC dataset (with ResNet-12 backbone). Green: actinic keratosis; yellow: vascular lesion; and purple: dermatofibroma.

model has learned solid knowledge about the rare diseases to generate consistent, meaningful pseudo labels rather than random ones.

For an intuitive perception of the learned few-shot representations, we use t-SNE (Van der Maaten and Hinton, 2008) to visualize the 5-shot features of all query-set rare disease samples of a specific FSL task (Fig. 6). The features of the best-performing method (besides ours) in each genre of FSL methods are visualized, too, including DAML (supervised meta-learning), UMTRA (unsupervised meta-learning), SRL-distill (supervised representation learning), and MoCo_v1 (unsupervised representation learning). As we can see, the dots of different colors mostly

Table 10

Evaluation results (in %) and comparison with SOTA FSL methods on the ISIC dataset with SENet (Hu et al., 2018) as the backbone. Standard deviation of the accuracy is parenthesized. A_1 , A_2 , and A_3 correspond to dermatofibroma, actinic keratosis, and vascular lesion, respectively.

Method	$(N, K) = (3, 1)$					$(N, K) = (3, 3)$					$(N, K) = (3, 5)$				
	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg	Accuracy	A_1	A_2	A_3	Avg
Training from scratch	37.80* (3.39)	33.10	42.51	30.70	36.03*	43.18* (4.23)	37.92	48.10	35.95	41.29*	49.37* (3.72)	43.32	50.11	52.49	48.82*
<i>Supervised Meta-learning (SML)</i>															
MAML	47.43* (3.23)	41.75	50.61	44.68	46.12*	53.86* (2.08)	55.26	53.06	54.60	54.20*	58.77* (2.35)	56.25	56.48	66.18	59.42*
RelationNet	39.69* (2.39)	34.71	42.56	37.10	38.52*	48.81* (1.71)	31.98	51.53	56.02	47.08*	51.95* (1.60)	36.25	54.93	57.55	50.17*
ProtoNet	36.54* (3.66)	29.50	37.29	40.51	35.96*	39.72* (4.61)	42.92	42.91	29.70	38.81*	59.54* (2.98)	62.28	59.20	58.15	59.79*
DAML	37.52* (3.16)	27.19	38.34	43.97	36.75*	55.48* (3.62)	52.67	54.01	61.15	55.83*	63.12* (3.11)	63.39	62.34	64.74	63.40*
<i>Unsupervised Meta-learning (UML)</i>															
UMTRA	46.46* (1.46)	47.36	47.23	43.97	46.26*	58.78* (5.91)	56.25	56.48	66.18	59.42*	63.79* (5.32)	69.09	62.11	63.50	64.62*
CACTUs-MAML	53.87* (1.77)	55.26	53.06	54.60	54.20*	59.13* (1.58)	62.50	56.48	62.58	60.17*	65.02* (3.95)	61.81	67.70	61.31	63.96*
CACTUs-ProtoNets	57.14* (5.67)	54.38	57.97	57.44	56.73*	63.65* (3.96)	66.96	62.34	64.02	64.24*	68.19* (2.27)	67.27	66.77	72.26	68.62*
<i>Supervised Representation Learning (SRL)</i>															
SRL-simple	62.99* (2.91)	66.66	63.19	59.57	63.10*	68.52* (3.64)	63.39	71.60	65.46	67.24*	71.00* (4.67)	69.09	72.04	70.07	70.55*
SRL-distill	64.71* (2.97)	65.78	61.34	71.63	65.86*	69.74* (3.62)	63.39	72.53	68.34	68.50*	72.05* (5.15)	70.90	73.60	69.34	71.47*
<i>Unsupervised Representation Learning (URL)</i>															
MoCo_v1	67.46 (4.92)	41.22	93.86	27.65	57.55	71.82 (5.25)	56.25	89.50	43.16	65.18	78.55* (1.90)	53.63	87.57	77.37	74.28*
Hbr.Dst.	68.32 (4.89)	64.03	86.19	30.49	62.26	70.44 (2.25)	69.17	76.20	58.02	68.45	80.49 (2.73)	50.00	89.75	83.21	75.86
Hbr.Dst.-DDL	67.12 (2.82)	61.40	80.42	41.00	62.49	72.17 (3.94)	57.14	90.74	41.00	65.26	79.43 (2.75)	45.45	88.50	85.40	74.70
Hbr.Dst.-DIC	68.76 (3.27)	62.05	74.84	60.13	66.45	72.47 (2.67)	68.51	79.00	60.44	70.11	81.19 (3.37)	57.27	90.99	77.37	76.71

* $p < 0.05$ for comparison with the proposed Hbr.Dst.-DIC.

Table 11

Performance (in %) comparison with the Oracle model (i.e., trained with hundreds of rare disease samples) on the ISIC dataset (ResNet-12 backbone). The best-performing method in each genre of FSL methods (SML: supervised meta-learning; UML: unsupervised meta-learning; SRL: supervised representation learning; and URL: unsupervised representation learning) is included, too (NA: not applicable). Note that our 1-shot results are superior to the 5-shot results of others, so we do not include their lower-shot results. Standard deviation of the accuracy is parenthesized.

Method	Group	Shot	Accuracy	A_1	A_2	A_3	Avg
Train scratch	NA	5	45.45* (1.12)	43.42	51.54	48.64	47.87*
DAML	SML	5	52.33* (2.17)	54.38	52.57	51.35	52.77*
UMTRA	UML	5	60.29* (3.79)	57.01	62.88	70.27	63.39*
SRL-distill	SRL	5	70.71* (5.08)	70.61	65.97	81.08	72.55*
		1	74.43* (3.35)	71.49	74.22	83.78	76.50*
Hbr.Dst.	URL	3	76.69* (2.98)	74.12	80.41	79.72	78.08
-DIC (ours)		5	80.70 (3.17)	77.78	87.18	74.28	79.75
		20	84.22 (2.01)	82.50	87.82	85.08	84.85
Oracle	NA	All	88.88	92.59	85.48	92.85	90.31

* $p < 0.05$ for comparison with our 5-shot results.

mingle together for “Train Scratch”, DAML, and UMTRA, whereas SRL-distill modestly improves the separability of the green and yellow dots. In contrast, MoCo_v1, our baseline model in this work, greatly improves the separability of the green and yellow dots (almost linearly separable) but still cannot handle the purple dots well. Eventually, our proposed Hbr.Dst.-DIC exhibits the best and most balanced separability for all three colors. These visualizations agree with the quantitative evaluation results presented in Table 1 and demonstrate that the few-shot representation learned by our method is superior in differentiating the rare diseases.

4.5.3. Comparison with Oracle model

Next, we compare the performance of our final FSL models trained with pseudo labels and conservative learning on the CDNC data with that of an Oracle model trained with real labels on substantially more rare disease data than the few-shot settings. Concretely, we split all samples of actinic keratosis (327 images), dermatofibroma (115 images), and vascular lesion (142 images) of the ISIC dataset (recall that these diseases are the simulated “rare” diseases in Tables 1 and 2) in the ratio of 7:1:2 for training, validation, and testing, respectively. It is worth mentioning that the testing set here is a subset of the query sets in the few-shot experiments, so the FSL models in Table 1 can be directly evaluated here without data leaks. Then, a supervised

ResNet-12 model is trained on the training set with the real labels. The results are shown in Table 11, together with other representative FSL methods. As we can see, our 5-shot results are the closest to the Oracle (total accuracy 80.70% versus 88.88%) among the FSL methods, outperforming others’ 5-shot results by at least 10% in total accuracy. Meanwhile, our 1-shot results are better than the 5-shot results of others. When we increase the shot number to 20, our method’s total accuracy increases to 84.22%, narrowing the gap from the Oracle to ~4%, yet with about six times fewer rare disease samples for training. These results suggest that the pseudo rare disease labels on CDNC instances can train the network effectively, asymptotically approaching the level of training with a few hundred rare disease samples and labels. Notwithstanding, the remaining gap of ~4% in total accuracy and the more appreciable gap in average accuracy (~5%) are likely due to the limited diversity of the support sets in FSL, suggesting the necessity for future research.

4.5.4. Impact of backbone network

In the previous section, we mainly experiment with two backbone networks (the ResNet-12 and 4-conv-block) for their prevalent use in the few-shot literature (e.g., Finn et al., 2017; Nichol and Schulman, 2018, to name a few). The results in Tables 1 and 2 (and Tables 3 and 5) show that the choice of the backbone network exerts a general impact on the performance of all compared methods. Intuitively, the




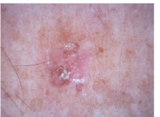
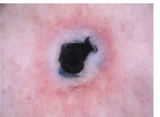
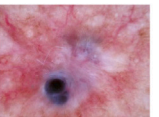
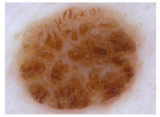
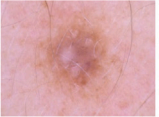
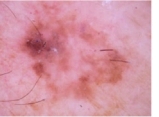
	(a)		(b)		(c)	
Error						
Pred	actinic keratosis	actinic keratosis	dermatofibroma	dermatofibroma	actinic keratosis	actinic keratosis
GT	dermatofibroma	dermatofibroma	actinic keratosis	actinic keratosis	vascular lesion	vascular lesion
Ref.						
GT	actinic keratosis		dermatofibroma		actinic keratosis	

Fig. 7. Examples of incorrectly classified samples by Hbr.Dst-DIC for the 5-shot setting on the ISIC dataset. “Error” shows incorrectly classified samples, while “Ref.” shows images visually similar to the error ones but belonging to the misclassified categories. Pred: model-predicted categories; and GT: ground truth categories.

more powerful the backbone, the better the performance. However, the trends of performance comparison are the same for both backbones, with our proposed Hbr.Dst-DIC achieving the highest total and average accuracies in all three shot settings. A natural question is whether the superior performance of our framework would persist if a different backbone network were used. To answer this question, we repeat the experiments in Table 1 (Table 2) with SENet (Hu et al., 2018) as the backbone. The SENet features the “squeeze-and-excitation” (SE) block that adaptively recalibrates channel-wise feature responses by explicitly modeling channel interdependencies. In this experiment, the ResNet-50 (He et al., 2020) version of SENet is used. The results in Table 10 show similar trends to Table 1 (hence, we do not repeat the description here), and Hbr.Dst-DIC again achieves the best performance in all shot settings. These results demonstrate our framework’s effectiveness on varying backbones and consistent superiority to compared methods.

4.5.5. Qualitative failure analysis

Fig. 7 shows the proposed method’s typical misclassification examples on the ISIC dataset. It can be seen that the misclassified samples are visually similar to some reference images belonging to the categories in which they are misclassified. For example, the white balance of the images in group (a) seems to be similar to each other but different from images in other groups, in addition to their similar central patterns. Hence, introducing more rigorous color/hue augmentation might strengthen the performance of our method. The images in group (c) are misclassified probably due to the reddish rings similar to the reference image despite the central difference of dark spots, highlighting the complex nature of the FSL rare disease classification problem.

5. Discussion and conclusion

In this work, we proposed a novel approach to rare disease imaging phenotype classification in two steps. First, we built a baseline model on unsupervised representation learning for a simple and label-free (with respect to the base dataset of common diseases and normal controls) framework, which achieved superior performance to existing FSL methods. Second, we further proposed to utilize the baseline model as the teacher model for a hybrid self-distillation integrating unsupervised contrastive learning and pseudo-label supervised classification. Also, we proposed dispersion-aware imbalance correction (DIC) during the distillation to alleviate the severe performance imbalance between the rare diseases. Experimental results suggested that the hybrid approach with DIC could effectively inject knowledge about the rare diseases into the representation learning process through the pseudo-labels, relieve overfitting to noise and bias in the small support set thanks to the contrastive learning, and reduce inter-class performance imbalance by focusing more on less well represented classes. Consequently, our new

approach set a new state of the art (SOTA) for rare disease imaging phenotype classification.

Computer-aided diagnosis (CAD) has achieved remarkable success with the advance in deep learning techniques driven by large labeled datasets. However, automatic rare disease imaging phenotype classification still faces major challenges due to the scarcity of training samples. Existing few-shot learning (FSL) approaches (Chen et al., 2022; Li et al., 2020; Paul et al., 2021) relied on a large, labeled base dataset of common diseases and normal controls (CDNC) to learn generalizable knowledge transferable to the rare diseases. In this work, we took a step further to eliminate the need for labeling the base dataset, which is often expertise-demanding, time-consuming, and laborious. Concretely, our work proposed unsupervised representation learning on unlabeled CDNC data to aid rare disease diagnosis, and, to the best of our knowledge, was the first of its kind. To this end, our approach broadened the practical application scenarios compared to existing ones, by not requiring the auxiliary base dataset to be labeled. Meanwhile, our approach achieved this without sacrificing performance—in fact, as our experiments showed, it outperformed a variety of up-to-date FSL approaches that did require a labeled base dataset.

In the comparison to SOTA approaches (Section 4.3), we compared four groups of methods: supervised meta-learning, unsupervised meta-learning, supervised representation learning, and unsupervised representation learning. The results showed that the representation learning methods generally outperformed the meta-learning ones in most settings. This suggested that sophisticated meta-learning algorithms were not a must for few-shot learning, and it was the feature reuse that was the main factor for fast adaptation (Raghu et al., 2019). Representation learning on the CDNC data prepared the model with excellent feature extraction ability, and then the rare disease classification could reuse the common features to make the model adapt fast. In addition, we empirically and surprisingly found that many unsupervised representation learning (URL) methods outperformed supervised meta-learning and representation learning in this work. A possible explanation may be that the limited number of classes in the base dataset D_{base} presented limited variations and made the supervised representations overfit to their differentiation, whereas the task of instance discrimination in unsupervised representation learning forced the networks to learn more diverse representations that were more generalizable on novel classes (Sun et al., 2022).

Moreover, the proposed hybrid distillation brought further performance improvement upon the baseline URL model. On the one hand, the results strongly supported our hypothesis that, by extracting the knowledge about the rare diseases from the small support set and injecting it into the representation learning process via pseudo-label supervision, we could exploit the large CDNC dataset to learn representations and classifiers that can better distinguish the rare diseases.

On the other hand, the ablation study results indicated that distilling with the pseudo-label supervision (by \mathcal{L}_{cls}) was not enough, and incorporating the self-supervising contrastive loss \mathcal{L}_{con} was indispensable to the superb performance of our full model (cf. row (b) and row (c) in Table 7). We conjecture this was because \mathcal{L}_{con} helped avoid overfitting to the support set of the rare diseases, which may be affected by noise and bias due to its small size. Hence, the hybrid distillation fully utilized the large base dataset as well as the small support set, meanwhile effectively controlling overfitting to potential noise and bias in the latter.

Last but not least, observing the severe performance imbalance between the rare diseases, we proposed a novel dispersion-aware imbalance correction (DIC). Unlike the commonly encountered performance imbalance which is often a result of imbalanced training classes, we speculate that the imbalance here was caused by the insufficient representation of the rare diseases due to the randomness and small size of the support data. Therefore, existing solutions that relied on the size distribution of the training classes like the distribution-aware diversity loss (DDL) (Wang et al., 2021) were not applicable. Instead, we resorted to the extent of feature dispersion as a measure of intra-class diversity and proposed to focus more on less diverse classes. Experimental results showed that while the DDL did not seem to work in our context, our DIC effectively reduced the performance imbalance between classes (e.g., see the last two rows in Table 1).

Clinically, diagnosing rare diseases is complex and usually done with several procedures instead of imaging alone. This work focused on imaging phenotype classification of rare diseases. Instead of reaching a final, definite diagnosis, it was intended to facilitate the clinical image reading process and contribute to the integral diagnosis pipeline. In addition, our choice of imaging modality was partly because imaging data was more accessible in sizeable amounts for developing deep learning methods than other modalities (e.g., laboratory results like blood tests) of medical data (Lee et al., 2022). However, our framework also has the potential to be adapted and applied to aid rare disease classification with other data modalities, given enough commonalities between typical and rare diseases, which is usually the case. Meanwhile, to prove our framework's usability for real clinical applications, clinical validation with more in-depth analyses is necessary for future work.

This work had limitations. An underlying assumption of the proposed approach was that the diseases in the large base dataset should share common features with the target rare diseases. Usually, the assumption is implicitly satisfied when the base and target datasets belong to the same disease group (e.g., both skin lesions) and the same type (e.g., both optical imaging). It is unlikely for the proposed approach to work as effectively in the unintended scenario where the base and target datasets are two different disease groups or image types (such as optical and magnetic resonance imaging). In future work, it would be interesting to quantify the commonalities between the base dataset and target rare diseases required to guarantee a reasonable performance (and investigate how to quantify them). However, with that being said, we do expect the collection of *massive, inclusive* base datasets comprising many different diseases and imaging modalities to strengthen our proposed approach further. Recent studies have shown that modern deep networks are capable of learning and predicting from a wide range of inputs and achieving encouraging performance on a wide range of tasks (e.g., Xu et al., 2023; Zhang et al., 2023), especially when the networks and datasets are both large and with effective designs to unify various task and data formats. We envision this direction for future work.

CRedit authorship contribution statement

Jinghan Sun: Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Dong Wei:** Conceptualization, Formal analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Liansheng Wang:** Funding acquisition, Supervision, Writing – review & editing. **Yefeng Zheng:** Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We use publicly available data and cite them properly in this work. Our code will be released along with publication of the paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2019YFE0113900).

References

- Boudiaf, M., Kervadec, H., Masud, Z.I., Piantanida, P., Ben Ayed, I., Dolz, J., 2021. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13979–13988.
- Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* 106, 249–259.
- Byrd, J., Lipton, Z., 2019. What is the effect of importance weighting in deep learning? In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 872–881.
- Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 1567–1578.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *J. Artificial Intelligence Res.* 16, 321–357.
- Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Chen, Y., Guo, X., Xia, Y., Yuan, Y., 2022. Disentangle then calibrate: Selective treasure sharing for generalized rare disease diagnosis. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 512–522.
- Chen, X., He, K., 2021. Exploring simple Siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758.
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 1597–1607.
- Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C.F., Huang, J.-B., 2019. A closer look at few-shot classification. In: International Conference on Learning Representations.
- Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X., 2021b. Meta-baseline: Exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9062–9071.
- Chen, X., Yao, L., Zhou, T., Dong, J., Zhang, Y., 2021a. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. *Pattern Recognit.* 113, 107826.
- Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kallou, A., Liopyris, K., Marchetti, M., et al., 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*.
- Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S., 2019. Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9268–9277.
- Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 1126–1135.
- Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A., 2018. Born again neural networks. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 1607–1616.
- Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V., Zoph, B., 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2918–2928.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent—A new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* 33, 21271–21284.
- Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vol. 2, IEEE, pp. 1735–1742.

- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.
- Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. In: NeurIPS Deep Learning and Representation Learning Workshop.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9 (8), 1735–1780.
- Hsu, K., Levine, S., Finn, C., 2018. Unsupervised learning via meta-learning. In: International Conference on Learning Representations.
- Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.
- Huang, L., Zhang, C., Zhang, H., 2020. Self-adaptive training: Beyond empirical risk minimization. *arXiv preprint arXiv:2002.10319*.
- Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B., 2005. Pap-smear benchmark data for pattern classification. In: Proceedings of Annual Symposium on Nature Inspired Smart Information Systems. pp. 1–9.
- Jiang, X., Ding, L., Havaei, M., Jesson, A., Matwin, S., 2019. Task adaptive metric space for medium-shot medical image classification. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 147–155.
- Ker, J., Wang, L., Rao, J., Lim, T., 2018. Deep learning applications in medical image analysis. *IEEE Access* 6, 9375–9389.
- Khodadadeh, S., Bölöni, L., Shah, M., 2019. Unsupervised meta-learning for few-shot image classification. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 10132–10142.
- Kim, J., Jeong, J., Shin, J., 2020a. M2m: Imbalanced classification via major-to-minor translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13896–13905.
- Kim, K., Ji, B., Yoon, D., Hwang, S., 2020b. Self-knowledge distillation: A simple way for better generalization. *arXiv preprint arXiv:2006.12000*.
- Lee, J., Liu, C., Kim, J., Chen, Z., Sun, Y., Rogers, J.R., Chung, W.K., Weng, C., 2022. Deep learning for rare disease: A scoping review. *J. Biomed. Inform.* 104227.
- Lee, K., Maji, S., Ravichandran, A., Soatto, S., 2019. Meta-learning with differentiable convex optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10657–10665.
- Li, F.-F., Fergus, R., Perona, P., 2006. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* 28 (4), 594–611.
- Li, X., Yu, L., Jin, Y., Fu, C.-W., Xing, L., Heng, P.-A., 2020. Difficulty-aware meta-learning for rare disease diagnosis. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 357–366.
- Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88.
- Liu, S., Han, J., Puyal, E.L., Kontaxis, S., Sun, S., Locatelli, P., Dineley, J., Pokorny, F.B., Dalla Costa, G., Leocani, L., et al., 2022. Fitbeat: COVID-19 estimation based on wristband heart rate using a contrastive convolutional auto-encoder. *Pattern Recognit.* 123, 108403.
- Liu, J., Sun, Y., Han, C., Dou, Z., Li, W., 2020. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2970–2979.
- Mobahi, H., Farajtabar, M., Bartlett, P., 2020. Self-distillation amplifies regularization in Hilbert space. *Adv. Neural Inf. Process. Syst.* 33, 3351–3361.
- Murthy, R., Haji, S., Sambhav, K., Grover, S., Chalam, K., 2016. Clinical applications of spectral domain optical coherence tomography in retinal diseases. *Biomed. J.* 39 (2), 107–120.
- Nguangwakap, S., Lambert, D.M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., Rath, A., 2020. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* 28 (2), 165–173.
- Nichol, A., Schulman, J., 2018. Reptile: A scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* 2, 4.
- Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Paul, A., Tang, Y.-X., Shen, T.C., Summers, R.M., 2021. Discriminative ensemble learning for few-shot chest X-ray diagnosis. *Med. Image Anal.* 68, 101911.
- Raghu, A., Raghu, M., Bengio, S., Vinyals, O., 2019. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. *arXiv preprint arXiv:1909.09157*.
- Ravichandran, A., Bhotika, R., Soatto, S., 2019. Few-shot learning with embedded class models and shot-free meta training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 331–339.
- Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19 (1), 221–248.
- Shi, X., Wei, D., Zhang, Y., Lu, D., Ning, M., Chen, J., Ma, K., Zheng, Y., 2022. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 151–168.
- Shorffuzzaman, M., Hossain, M.S., 2021. MetaCOVID: A siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. *Pattern Recognit.* 113, 107700.
- Singh, R., Bharti, V., Purohit, V., Kumar, A., Singh, A.K., Singh, S.K., 2021. MetaMed: Few-shot medical image classification using gradient-based meta-learning. *Pattern Recognit.* 120, 108111.
- Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4080–4090.
- Steiner, B., DeVito, Z., Chintala, S., Gross, S., Paszke, A., Massa, F., Lerer, A., Chanan, G., Lin, Z., Yang, E., et al., 2019. PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32, 8026–8037.
- Stolk, P., Willemen, M.J., Leufkens, H.G., 2006. Rare essentials: Drugs for rare diseases as essential medicines. *Bull. World Health Org.* 84, 745–751.
- Sun, J., Wei, D., Ma, K., Wang, L., Zheng, Y., 2021. Unsupervised representation learning meets pseudo-label supervised self-distillation: A new approach to rare disease classification. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 519–529.
- Sun, J., Wei, D., Ma, K., Wang, L., Zheng, Y., 2022. Boost supervised pretraining for visual transfer learning: Implications of self-supervised contrastive representation learning. In: Association for the Advancement of Artificial Intelligence.
- Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M., 2018. Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q., 2021. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1685–1694.
- Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J.B., Isola, P., 2020. Rethinking few-shot image classification: a good embedding is all you need? In: Proceedings of the European Conference on Computer Vision. Springer, pp. 266–282.
- Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* 5 (1), 1–9.
- Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9 (11).
- Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D., 2016. Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. pp. 3637–3645.
- Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S., 2021. Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations.
- Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F., 2021. CREST: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10857–10866.
- Wu, N., Yu, L., Yang, X., Cheng, K.-T., Yan, Z., 2023. FedIIC: Towards robust federated learning for class-imbalanced medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 692–702.
- Xu, L., Ni, Z., Liu, X., Wang, X., Li, H., Zhang, S., 2023. Learning a multi-task transformer via unified and customized instruction tuning for chest radiograph interpretation. *arXiv preprint arXiv:2311.01092*.
- Yoo, T.K., Choi, J.Y., Kim, H.K., 2021. Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. *Med. Biol. Eng. Comput.* 59, 401–415.
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S., 2021. Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. Proceedings of Machine Learning Research, pp. 12310–12320.
- Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., Yue, X., 2023. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*.
- Zhang, D., Jin, M., Cao, P., 2020. ST-MetaDiagnosis: Meta learning with spatial transform for rare skin disease diagnosis. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 2153–2160.
- Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K., 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3713–3722.
- Zhu, W., Liao, H., Li, W., Li, W., Luo, J., 2020. Alleviating the incompatibility between cross entropy loss and episode training for few-shot skin disease classification. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 330–339.