Dynamic Subcluster-Aware Network for Few-Shot Skin Disease Classification

Shuhan Li, Xiaomeng Li, *Member, IEEE*, Xiaowei Xu, and Kwang-Ting Cheng, *Fellow, IEEE*

***Abstract*— This article addresses the problem of few-shot skin disease classification by introducing a novel approach called the subcluster-aware network (SCAN) that enhances accuracy in diagnosing rare skin diseases. The key insight motivating the design of SCAN is the observation that skin disease images within a class often exhibit multiple subclusters, characterized by distinct variations in appearance. To improve the performance of few-shot learning (FSL), we focus on learning a high-quality feature encoder that captures the unique subclustered representations within each disease class, enabling better characterization of feature distributions. Specifically, SCAN follows a dual-branch framework, where the first branch learns classwise features to distinguish different skin diseases, and the second branch aims to learn features, which can effectively partition each class into several groups so as to preserve the subclustered structure within each class. To achieve the objective of the second branch, we present a cluster loss to learn image similarities via unsupervised clustering. To ensure that the samples in each subcluster are from the same class, we further design a purity loss to refine the unsupervised clustering results. We evaluate the proposed approach on two public datasets for few-shot skin disease classification. The experimental results validate that our framework outperforms the state-of-the-art methods by around 2%–5% in terms of sensitivity, specificity, accuracy, and F1-score on the SD-198 and Derm7pt datasets.**

***Index Terms*— Few-shot learning (FSL), rare skin disease classification.**

I. Introduction

IN THE past decade, advances in deep convolutional neural networks and the availability of a large number of annotated images have continued to push the boundaries in a variety of medical image analysis tasks, such as organ segmentation [1], [2], tumor segmentation [3], [4], [5], [6], and disease screening [7], [8], [9], [10], [11]. Apart from the relatively sufficient examples of the common diseases, there are more than 6000 known rare diseases, at much lower prevalence, affecting 7% of the population worldwide [12]. The diagnosis of rare conditions is challenging for clinicians due to a lack of experience and for machines due to a lack of clinical samples. Rare skin disease diagnosis is one of the application domains that suffer from data deficiency. For example, the SD-198 dataset [13] consists of 6584 skin images with 198 disease classes, and 70 of these classes contain relatively fewer images than the other skin disease classes, as shown in Fig. 1. This article aims to develop a deep learning model, which is trained with images of common skin diseases while generalizing well on rare skin diseases.

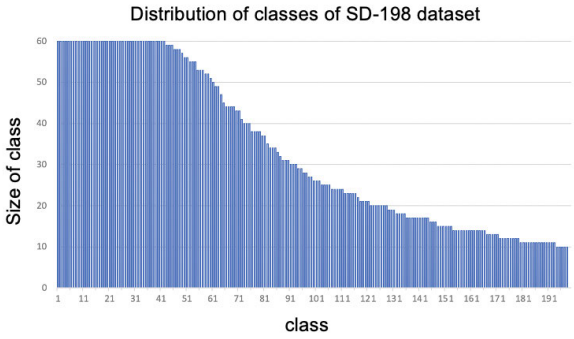

Fig. 1. Long-tailed class distribution in the SD-198 dataset. The X-axis refers to 198 skin disease classes, and the Y-axis refers to the number of images in the corresponding class.

To address the challenge of rare skin disease diagnosis, few-shot learning (FSL) has emerged as a promising solution. FSL aims to learn transferable knowledge on common classes (base classes) and apply it to the rare ones (novel classes) with only a handful of labeled data. Current FSL methods for few-shot skin disease classification can be broadly divided into two groups: meta-learning-based [14], [15], [16], [17], [18], [19], [20], [21] and transfer-learning-based [22], [23], [24], [25], [26], [27]. Meta-learning-based methods are mainly based on model-agnostic meta-learning (MAML) [28] and Prototypical Networks [29], with a focus on different problems in rare skin disease classification. For example, Li et al. [15] built a difficulty-aware meta-learning (DAML) model based on MAML [28] to address the different weights of sampled tasks in the imbalanced skin dataset. The other group of FSL methods, transfer-learning-based methods, aim to learn a powerful feature extractor in the pretraining stage with sufficient data, which can be quickly adapted to novel classes in the fine-tuning stage with a few novel examples. For example, Xiao et al. [27] designed a multitask framework that leverages contrastive learning as the auxiliary task to improve the performance of the classification task.

Manuscript received 15 July 2022; revised 28 August 2023 and 8 November 2023; accepted 22 November 2023. This work was supported in part by the Foshan HKUST Projects under Grant FSUST21-HKUST10E and Grant FSUST21-HKUST11E, and in part by the Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under Grant HZQB-KCZYB-2020083. *(Corresponding authors: Xiaomeng Li; Xiaowei Xu.)*

Shuhan Li is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China.

Xiaomeng Li is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China, and also with HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen (e-mail: eexmli@ust.hk).

Xiaowei Xu is with the Department of Cardiovascular Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University, Guangzhou 510080, China.

Kwang-Ting Cheng is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China (e-mail: timcheng@ust.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3336765.

Digital Object Identifier 10.1109/TNNLS.2023.3336765

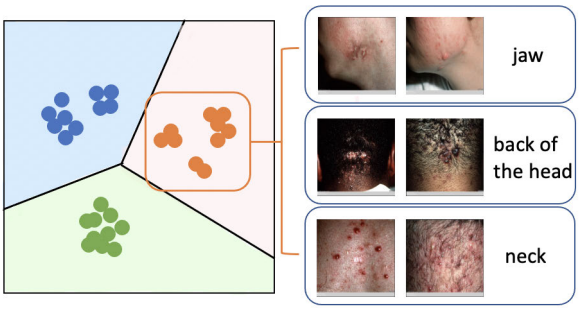

Fig. 2. Left: example of features learned by our framework. Three classes are separated by their class labels (shown in different colors), and various subgroups are generated automatically for each class. Right: example of a class forming three intrinsic subgroups, with different appearances in different body parts.

However, there exists a key challenge for rare skin disease classification that has rarely been addressed by the previous FSL methods. Skin disease images have large intraclass variations [30] and form multiple latent groups within a class, i.e., subclusters, for some classes, as shown in Fig. 2. To address this issue, the prototypical clustering networks (PCNs) [14] were proposed to learn multiple prototypes for each class. In their experiments, they manually set a fixed number of clusters for all classes, typically dividing each class into two subclusters. This approach may lead to inaccurate structure learning for certain classes, as different classes often display varying subcluster structures. Instead of a fixed subcluster number for all classes, we should identify various subcluster numbers for different classes.

Therefore, in this article, we design a novel dual-branch subcluster-aware network (SCAN), for few-shot skin disease classification. Our approach stands out from previous methods by dynamically identifying the unique subclustered structures within each class. This innovation is inspired by the discovery of intrinsic subclustered structures in skin disease datasets. By learning these structures from skin images, the network is able to extract more precise feature representations on the base classes, leading to enhanced performance on novel classes. To achieve this objective, SCAN incorporates a powerful feature encoder that consists of two branches. The class branch learns classwise features to distinguish different skin diseases, while the cluster branch learns subclustered features to preserve the subclustered structure within each class. To facilitate the learning of dynamic subclustered representation, we introduce a cluster branch with a cluster loss, which employs unsupervised clustering to partition all base class samples into several clusters. To ensure that the samples in each subcluster are from the same class, we further design a purity loss to refine the unsupervised clustering results. Inspired by the concept of triplet loss, the purity loss guides the anchor away from the wrong cluster center and pulls it closer to the correct class center. Notably, our framework can identify the different numbers of subclusters according to the different class structures. In the fine-tuning stage, we evaluate the performance of the pretrained encoder on novel classes with sampled episodes.

Experimental results on two skin disease benchmark datasets, SD-198 [13] and Derm7pt [31], show that our method outperforms the state-of-the-art methods in both one-shot and five-shot settings. Our method achieves significant improvements compared to the prior best method in various metrics such as sensitivity, specificity, accuracy, and F1-score, with approximately 2% improvement observed for the SD-198 dataset and 5% improvement for the Derm7pt dataset. In summary, our contributions can be summarized as follows.

1) We present a dual-branch framework for rare skin disease classification following the paradigm of transfer-learning-based FSL methods. Our framework effectively identifies inherent and unique subcluster structures within each common disease class, leading to improved classification accuracy for rare disease classes.
2) Within our framework, we develop a powerful feature encoder that combines supervised classification (class branch) and unsupervised clustering (cluster branch). Moreover, we introduce the purity loss to improve the result of intraclass clustering in the unsupervised cluster branch.
3) Our method is evaluated on two public skin disease datasets, SD-198 [13] and Derm7pt [31], and achieves state-of-the-art results for few-shot skin disease classification. The code is available at https://github.com/xmed-lab/SCAN.

II. Related Work

A. FSL in Computer Vision

FSL aims to acquire prior knowledge from base classes (with large labeled examples) and propagate it to novel classes (with insufficient labeled examples, unseen during the training stage). FSL approaches can be categorized into two groups based on their training processes: meta-learning- and transfer-learning-based methods.

Meta-learning-based methods, which are motivated by meta-learning principles, have been extensively explored in FSL [28], [29], [32], [33], [34], [35], [36], [37], [38]. Early works, such as MAML [28], Reptile [32], Matching Networks [33], Prototypical Networks [29], and Relation Networks [34], adopt a meta-learning framework that trains the model episodically, enabling it to learn how to update the network using a few sampled examples. This enables these methods to achieve promising accuracy on novel classes with very limited labeled data. Meta-learning-based methods can be further divided into two subbranches. The first subbranch is gradient-based methods, such as MAML and Reptile, which aim to learn good initialization parameters for each episode. These methods can quickly adapt to unseen classes within one or a few gradient descent steps. The second subbranch is metric-based methods, which utilize encoded feature vectors and employ a distance metric to assign labels based on the nearest neighbor principle. These methods use similarity measures such as cosine similarity (e.g., Matching Networks) or Euclidean distance (e.g., Prototypical Networks).

In recent years, a second group of FSL methods based on transfer learning has emerged [39], [40], [41], [42], [43], [44], [45], [46], [47], [48]. These methods, including Baseline [39], S2M2_R [43], and PT-MAP [46], train a standard classification network on base classes and fine-tune the classifier head on episodes generated from novel classes. As learning

meaningful feature representations is essential to various deep-learning tasks [49], [50], [51], [52], the transfer-learning-based methods aim to train a strong feature extractor that generates transferable features for the novel set. Experimental results have shown that simple transfer-learning-based methods, such as training a classification network with standard cross-entropy loss, can achieve comparable performance to previous FSL methods while offering a simpler and more effective process [39], [53]. Due to their superior performance of transfer-learning-based methods, we explore them as ways to predict rare skin diseases.

B. FSL for Skin Disease Classification

Due to the unbalanced distribution of skin disease classes and the deficiency of images for rare conditions, it is essential to apply FSL methods to alleviate the reliance on massive training data. Previous studies have explored meta-learning-based methods, particularly within the gradient- and metric-based subbranches. In the gradient-based subbranch, Li et al. [15] built a DAML method, which is based on the MAML framework. DAML considers the varying difficulties of different episodic tasks. The model adjusts the weights of task-specific losses by downweighting the easy tasks and increasing the weights of the hard tasks. This approach reduces the importance of easier tasks and emphasizes the significance of more challenging tasks. Singh et al. [20] trained the Reptile [32] model with additional regularization techniques such as mixup, cutout, and cutmix. These techniques enhance model generalization by introducing variations and augmentations during training. The incorporation of these regularization methods improved the model's performance by 2%–5%. In the metric-based subbranch, more methods are proposed. Mahajan et al. [16] replaced the conventional convolutional layers in Reptile and Prototypical Networks with group equivariant convolutions [54]. This modification aims to capture invariant features after undergoing various transformations. Skin diseases typically lack a dominant global orientation structure, thus making group equivariant convolutions suitable for extracting informative features. Zhu et al. [18] identified an incompatibility issue between cross-entropy loss and episode training and proposed the query-relative loss. This novel loss function further enhances the behavior of metric-based methods. In addition, Zhu et al. [19] extended the Prototypical Networks framework by introducing different temperatures for different categories. By penalizing query samples that are not sufficiently close to their corresponding categories, this approach strengthens the generalization ability of the learned metric. For the category of transfer-learning-based algorithms, Xiao et al. [27] designed a multitask framework that leverages contrastive learning as the auxiliary task to enhance the performance of the few-shot classification task. Dai et al. [26] proposed a dual-encoder architecture to combine the knowledge from a large-scale image dataset and a few-shot medical image dataset. Some other works [23], [24], [25] utilized unlabeled data and self-supervised learning to mitigate the performance degradation of few-shot learners. In our work, we do not use additional data from other domains or unlabeled data. We focus on training the effective feature encoder on common skin disease images and transferring that knowledge to rare skin disease images.

In addition, few methods consider the subcluster structures within skin disease classes, except for PCNs proposed by Prabhu et al. [14]. The authors extend Prototypical Networks by representing each class as multiple prototypes, instead of a single one. However, PCN divides each class into fixed k subclusters, which is determined without considering the specific clustering results for different classes. In contrast to PCN, our designed framework aims to dynamically learn the specific intrinsic subclusters for each skin disease type. By doing so, we can generate more accurate feature representations for transference to unseen classes. Our approach goes beyond the fixed k subcluster division in PCN and allows for a more flexible and tailored representation learning process.

III. Methodology

A. Problem Setting

In FSL, data are partitioned into two disjoint sets: the base set and the novel set. We denote the images from the base set as $D_b = \{(x_i, y_i)\}_{i=1}^{N_b}$, which contains N_b labeled images from C_b base classes and $y_i \in \{1, \ldots, C_b\}$. Let $D_n = \{(x_i, y_i)\}_{i=1}^{N_n}$ denote the images from the novel set, with few samples from C_n novel classes, where $C_b \cap C_n = \emptyset$. The main goal of few-shot skin disease classification is to learn rich and transferable feature representations from the base set such that they can easily be adapted to the novel set with only a few labeled data.

To evaluate the generalization ability on the unseen classes, we sample N-way K-shot episodes from novel set D_n following the most common way proposed by Vinyals et al. [33]. Each episode contains a support set for fast adaptation and a query set for model evaluation. In the support set, N classes are randomly selected from C_n novel classes and K labeled images are sampled from each class, denoted as $S = \{(x_i, y_i)\}_{i=1}^{N\times K}$. In the query set, q images are picked from the same N classes without their labels, denoted as $Q = \{(x_i)\}_{i=1}^{N\times q}$. The label information of query set Q is only available when computing the accuracy.

Our work is based on the transfer-learning FSL paradigm. Therefore, we train a feature encoder in the pretraining stage with the data from the base set, as shown in Fig. 3. Then, we preserve the feature extractor f_θ of the trained model and freeze the network parameters. In the following fine-tuning stage, we fine-tune a new classifier by using the images from the support set in a sampled episode and evaluate the accuracy on the query set, as shown in Fig. 4.

B. Framework Overview

To specifically identify the subclusters within each class, we design the SCAN in the pretraining stage. The overall framework is shown in Fig. 3, which consists of a dual-branch network structure as well as several auxiliary memory banks.

The dual-branch network includes a feature extractor f_θ for feature encoding, a projection head z_θ for dimensional reduction, and two linear classifiers $C(.|\mathbf{W}^{\text{base}}_{\text{class}})$ and $C(.|\mathbf{W}^{\text{base}}_{\text{cluster}})$ for class and cluster prediction, respectively. We train the network using the labeled data in the base set by minimizing

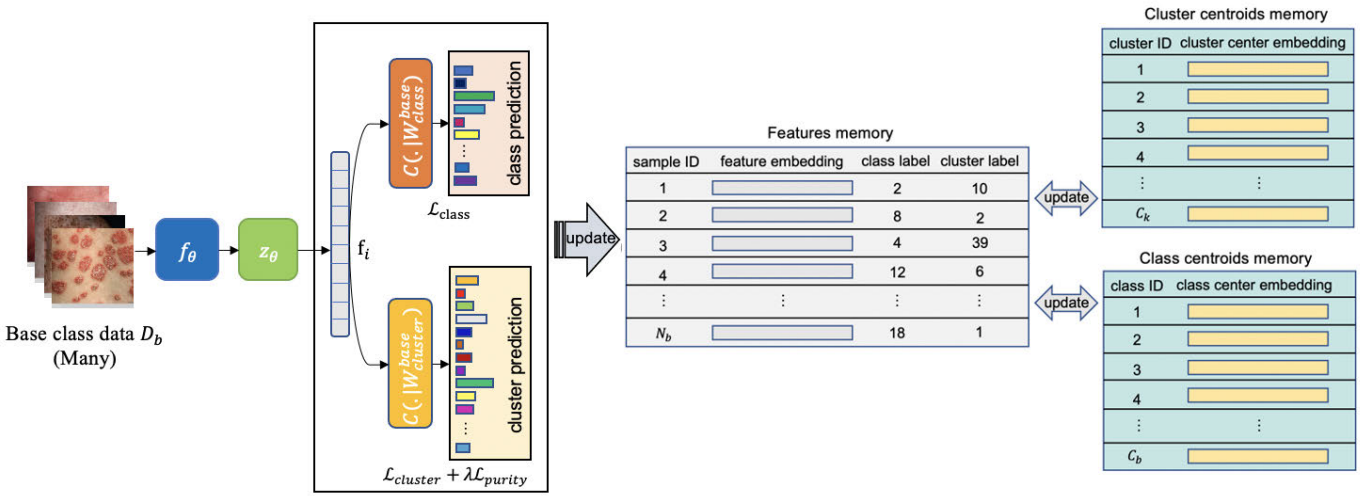

Fig. 3. Overview of the proposed SCAN. We train this network on base class data in the pretraining stage.

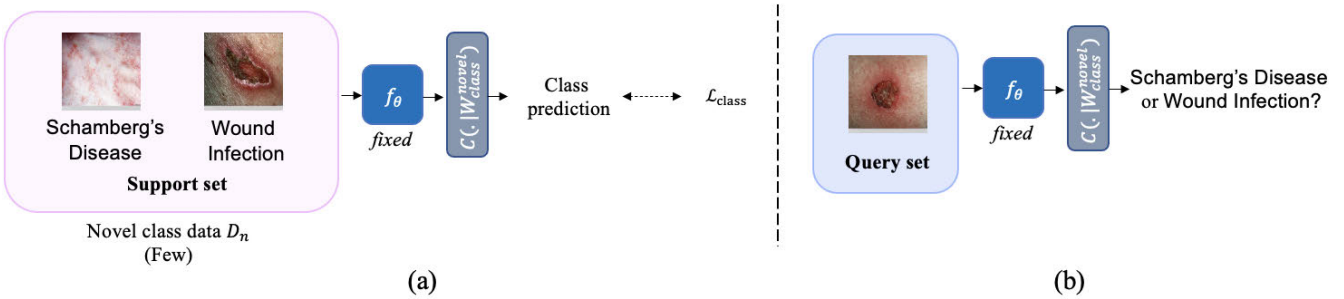

Fig. 4. Example of fine-tuning stage for a two-way one-shot task on novel class set. We utilize the feature encoder f_θ trained in the pretraining stage and fix the parameters. (a) In each episode, a classifier is trained by the support set images. (b) Evaluating on the query set images.

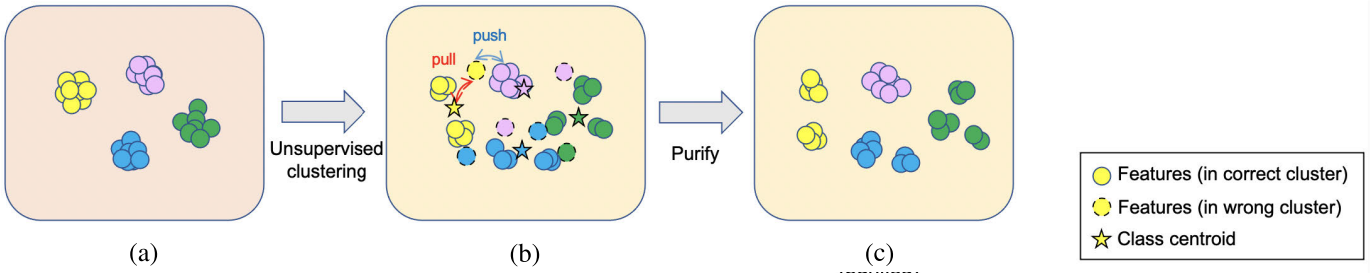

Fig. 5. Effects of cluster loss $\mathcal{L}_{\text{cluster}}$ [from (a) to (b)] and purity loss $\mathcal{L}_{\text{purity}}$ [from (b) to (c)]. (a) Class aware. (b) Subcluster aware (unsupervised). (c) Subcluster aware (purified).

three losses: $\mathcal{L}_{\text{class}}$, $\mathcal{L}_{\text{cluster}}$, and $\mathcal{L}_{\text{purity}}$. $\mathcal{L}_{\text{class}}$ is a standard cross-entropy loss computed for the class branch, which aims to classify skin disease according to the class labels. The other two losses are used in the cluster branch, with more details introduced in Section III-C.

C. Subcluster-Aware Network

Due to the intrinsic subcluster structure in the skin disease datasets, the model trained with the classwise cross-entropy loss cannot learn the structural relationships within each skin disease class, as shown in Fig. 5(a). Considering the subcluster structures for skin disease images, we present a novel SCAN to learn subcluster embeddings for skin disease classification.

To achieve this, we create three memory banks, i.e., feature memory, class centroid memory, and cluster centroid memory, as shown in Fig. 3. The feature memory bank saves the feature embeddings and their class and cluster labels. The size of the feature memory bank is set as N_b, which is the total number of training images. At the beginning of the training, we utilize the feature extractor f_θ to embed all the training images into feature embeddings and save them in the feature memory bank. Then, we initialize the cluster labels by applying K-means on all the saved feature embeddings. The class centroid memory saves the center embedding of each class by averaging all the features from the corresponding class. The size of the class centroid memory bank is set to the total number of classes in the base class dataset. Similar to the class centroid memory, the cluster centroid memory stores the mean feature of each cluster, and the size is the number of total clusters.

Then, in each forward propagation, the network and memory banks are updated through the procedures in the following.

1) Forward Propagation: We derive the feature embedding $\mathbf{f}_i = z_\theta(f_\theta(x_i)) \in \mathbb{R}^d$ for an input image x_i, where d refers to the dimension of $\mathbf{f}_i$. After sending $\mathbf{f}_i$ to two classifiers, we obtain the class probability prediction $p_i \in \mathbb{R}^{C_b}$ and the cluster probability prediction $p'_i \in \mathbb{R}^{C_k}$ correspondingly. Here, C_b is the number of base classes and C_k is the number of clusters. We set $C_k > C_b$ such that the model can acquire finer structures than the original granularity defined by the class labels.

2) Computing Class Loss $\mathcal{L}_{class}$: A standard cross-entropy loss in (1) is used to compute class loss $\mathcal{L}_{\text{class}}$

$$\mathcal{L}_{\text{class}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{c=1}^{C_b} -y_{i,c} \log p_{i,c} \tag{1}$$

$$y_{i,c} = \mathbb{1}[y_i = c] \tag{2}$$

where N_b denotes the amount of training data in base classes and $\mathbb{1}(\cdot)$ is the indicator function. $y_{i,c}$ is equal to 1 if c is the correct class label of input image x_i; otherwise, $y_{i,c}$ is 0. $p_{i,c}$ denotes the cth item in the class probability vector p_i, which represents the probability of x_i predicted as class c.

3) Computing Cluster Loss $\mathcal{L}_{cluster}$: With the cluster label of feature $\mathbf{f}_i$ recorded in the feature memory, denoted as y'_i, we calculate the cluster loss $\mathcal{L}_{\text{cluster}}$ by the following equation:

$$\mathcal{L}_{\text{cluster}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{s=1}^{C_k} -y'_{i,s} \log p'_{i,s} \tag{3}$$

$$y'_{i,s} = \mathbb{1}[y'_i = s]. \tag{4}$$

Similarly, $y'_{i,s}$ is equal to 1 if s is the cluster label of input image x_i, and otherwise, $y'_{i,s}$ is 0. $p'_{i,s}$ denotes the cluster probability of x_i predicted as cluster s.

4) Computing Purity Loss $\mathcal{L}_{purity}$: The above two losses, $\mathcal{L}_{\text{class}}$ and $\mathcal{L}_{\text{cluster}}$, are effective in forming the subcluster structure of the features. However, the clustering may not precisely occur within the class as the class labels are not utilized during cluster prediction. Samples from different classes are likely to be wrongly gathered into the same cluster because of the interclass similarity, as shown in Fig. 5(b). To address this issue, we propose a purity loss $\mathcal{L}_{\text{purity}}$ [see (5)] to refine the clustering results

$$\begin{aligned} \mathcal{L}_{\text{purity}} &= \sum_{i=1}^{N_b} \mathcal{L}_{\text{triplet}}\left(\mathbf{f}_i, \mathbf{f}_i^p, \mathbf{f}_i^n\right) \\ &= \sum_{i=1}^{N_b} \max\left(\left\|\mathbf{f}_i - \mathbf{f}_i^p\right\|_2^2 - \left\|\mathbf{f}_i - \mathbf{f}_i^n\right\|_2^2 + \alpha, 0\right) \end{aligned} \tag{5}$$

where $\mathbf{f}_i$ is the anchor, the positive sample $\mathbf{f}_i^p$ is the class center of the anchor's class, and the negative sample $\mathbf{f}_i^n$ is the nearest feature in the same cluster of $\mathbf{f}_i$ with a different class label. The margin between positive and negative pairs is denoted as α. The purity loss is defined as the sum of the triplet loss over all the samples in the base set.

The purity loss refines the clustering results by pulling the anchor toward its correct class and pushing it away from the wrong cluster, as shown in Fig. 5(c). Specifically, the loss is activated only when the distance between the anchor and positive item is longer than the distance between the anchor and negative sample by a margin α. By minimizing the purity loss, we can improve the quality of the learned subclusters and correct the improper clustering results caused by interclass similarity.

5) Updating Memory Banks: Finally, we update the contents in three memory banks according to the feature embeddings in the current epoch for the usage of next iteration.

In the feature memory bank, we smoothly update feature embeddings with a momentum decay rate $\beta \in (0, 1]$

$$\mathbf{f}_{m,i} \leftarrow \beta \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|} + (1 - \beta)\mathbf{f}_{m,i} \tag{6}$$

where $\mathbf{f}_i$ is the feature embedding with sample ID $i \in \{1, 2, \ldots, N_b\}$ and $\mathbf{f}_{m,i}$ represents the feature saved at the ith location in the feature memory bank. The cluster label is assigned to item i according to the nearest cluster center. The class labels remain the same since the ground truth does not change during training. In the class centroid memory bank, the centers of all the classes $c_i \in C_b$ are recalculated by the updated embeddings in the feature memory bank. In the cluster centroid memory bank, the centers of all the cluster $s_i \in C_k$ are recalculated by the updated embeddings and cluster labels.

To sum up, the whole loss function [see (7)] of our proposed SCAN framework is the summation of three parts

$$\mathcal{L}_{\text{SCAN}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{cluster}} + \lambda \mathcal{L}_{\text{purity}} \tag{7}$$

where λ is the weight for purity loss.

IV. Experiments

A. Datasets

SD-198 dataset [13] consists of 6584 clinical images captured by mobile phones or digital cameras. The dataset collects a diverse range of 198 skin disease classes, with each class containing 10–60 examples. The images differ in color, exposure, and illumination, and contain a wide range of patients of various ages, genders, skin colors, disease locations, and durations. To fairly compare with existing methods, we follow the settings of previous work [16]. In their study, a subset of images is selected from 90 out of 198 classes, and these images are subsequently divided into base and novel sets. The base set consists of 20 classes, with each class containing 60 images. The novel set consists of 70 classes, with the number of images per class ranging from 10 to 19. In total, the base set contains 1200 images and the novel set contains 965 images.

Derm7pt dataset [31] includes more than 2000 color images from 20 skin lesion classes. In line with previous research [16], we exclude two classes, namely, "miscellaneous" and "melanoma," from our analysis. Subsequently, the remaining 18 classes are divided into base and novel sets. According to their settings, the base set consists of 13 classes, with the number of images per class varying from 40 to 698. The novel set consists of five classes, with each class containing 10–34 images. The base set contains a total of 1892 images, while the novel set contains 114 images.

We resize all the input images to 80 × 80 pixels for the above two datasets and apply the standard augmentations,

TABLE I

TWO-WAY ONE-SHOT AND TWO-WAY TWO-SHOT FEW-SHOT CLASSIFICATION RESULTS (%) ON THE SD-198 DATASET. RESULTS OF OTHER METHODS ARE PRODUCED BY OUR REIMPLEMENTATION. †RESULTS REPORTED IN THIS ARTICLE. "-" REFERS TO NOT REPORTED IN THIS ARTICLE

Method	Backbone	2-way 1-shot		2-way 5-shot	
		Accuracy	F1-score	Accuracy	F1-score
PCN [14]	Conv4	70.03±1.42	70.78±1.61	84.95±1.15	85.87±1.12
SCAN (ours)		77.12±1.44	78.00±1.51	90.22±0.95	91.01±0.90
Meta-derm [16]	Conv6	65.3†	-	83.7†	-
SCAN (ours)		76.75±1.42	77.64±1.50	87.45±1.08	88.28±1.03
NCA [42]	WRN-28-10	71.27±1.50	71.27±1.50	83.30±1.20	84.23±1.19
Baseline [39]		75.72±1.47	76.64±1.56	88.95±1.00	89.66±0.97
S2M2_R [43]		76.42±1.52	77.51±1.59	90.32±0.89	90.97±0.89
NegMargin [44]		76.85±1.39	77.98±1.45	89.92±0.96	90.65±0.92
PT+NCM [46]		78.25±1.47	78.86±1.47	90.33±0.95	90.90±0.93
PEM_bE-NCM [47]		78.32±1.48	78.70±1.49	90.48±0.96	90.94±0.95
EASY [48]		78.80±1.50	79.44±1.51	90.87±0.98	91.43±0.96
SCAN (ours)		**80.20±1.44**	**81.21±1.46**	**91.48±0.88**	**92.08±0.85**

TABLE II

FIVE-WAY ONE-SHOT AND FIVE-WAY FIVE-SHOT FEW-SHOT CLASSIFICATION RESULTS (%) ON THE SD-198 DATASET. RESULTS OF OTHER METHODS ARE PRODUCED BY OUR REIMPLEMENTATION. †RESULTS REPORTED IN THIS ARTICLE. "-" REFERS TO NOT REPORTED IN THIS ARTICLE

Method	Backbone	5-way 1-shot				5-way 5-shot			
		Sensitivity	Specificity	Accuracy	F1-score	Sensitivity	Specificity	Accuracy	F1-score
PCN [14]	Conv4	44.79±0.96	86.20±0.24	77.92±0.38	45.59±1.03	64.78±0.99	91.19±0.25	85.91±0.39	65.70±1.02
SCAN (ours)		54.63±1.02	88.66±0.26	81.85±0.41	55.60±1.07	74.47±0.85	93.62±0.21	89.79±0.34	75.65±0.87
Meta-derm [16]	Conv6	-	-	-	-	-	-	-	-
SCAN (ours)		53.43±1.06	88.36±0.27	81.37±0.42	54.07±1.24	73.83±0.90	93.46±0.22	89.53±0.36	74.73±0.92
NCA [42]	WRN-28-10	45.43±1.02	86.36±0.26	78.17±0.41	45.91±1.08	61.83±0.96	90.46±0.24	84.73±0.38	62.83±1.01
Baseline [39]		51.57±1.05	87.89±0.26	80.63±0.42	52.54±1.11	73.65±0.94	93.41±0.24	89.46±0.38	74.71±0.96
S2M2_R [43]		54.79±1.07	88.70±0.27	81.91±0.43	55.49±1.13	77.11±0.83	94.28±0.21	90.84±0.33	78.17±0.84
NegMargin [44]		55.38±1.07	88.34±0.27	82.15±0.43	56.04±1.14	76.71±0.87	94.18±0.22	90.68±0.35	77.75±0.87
PT+NCM [46]		56.75±1.05	89.19±0.26	82.70±0.42	56.91±1.11	77.05±0.85	94.01±0.21	90.82±0.34	78.12±0.88
PEM_bE-NCM [47]		57.26±1.06	89.31±0.26	82.90±0.42	57.42±1.11	77.99±0.89	94.50±0.22	91.43±0.36	78.78±0.90
EASY [48]		57.55±1.05	89.39±0.26	82.99±0.43	57.77±1.12	78.81±0.88	94.71±0.21	91.65±0.35	79.53±0.89
SCAN (ours)		**58.08±1.09**	**89.52±0.27**	**83.23±0.44**	**58.75±1.14**	**80.41±0.78**	**95.10±0.19**	**92.16±0.31**	**81.43±0.77**

including random cropping, color jittering, rotation ($-30°$ to $+30°$), and horizontal flipping.

B. Implementation Details

To fairly compare with existing few-shot methods, we employ Conv4, Conv6 [29], and wide ResNet (WRN) [55] as backbones. To show the flexibility of our method, we further integrate our SCAN framework with ResNet18 and ResNet34 [56]. For the projection head z_θ, we use a nonlinear structure [fc-bn-relu-dropout-fc-relu], which reduces the feature embedding dimensions to 256. Both classifiers $C(.|\mathbf{W}^{\text{base}}_{\text{class}})$ and $C(.|\mathbf{W}^{\text{base}}_{\text{cluster}})$ have the same structure, i.e., a fully connected layer followed by a softmax function. The number of classes C_b is 20 for SD-198 and 13 for Derm7pt, while the number of clusters C_k is set to 40 and 25. For the hyperparameters in loss functions, we set α in (5) as 0.3, β in (6) as 0.5, and λ in (7) as 1. To train our SCAN model, we set the batch size to 16 and 64 for the SD-198 and Derm7pt datasets, respectively. We choose the SGD optimizer with a 0.0075 learning rate, 0.9 momentum, and 10^{-5} weight decay for 800 epochs of training.

To evaluate the performance on the novel set D_n, we freeze the parameters of the feature extractor f_θ and use it to encode images in a novel set to feature embeddings. Episodes are generated from the novel set randomly. We set two-way one-shot, two-way five-shot, five-way one-shot, and five-way five-shot for the SD-198 dataset, and set two-way one-shot and two-way five-shot for the Derm7pt dataset. Five query images are selected in each episode for both two datasets. For each episode, we train an episodic-specific linear classifier on the support set with the augmentation techniques introduced in [45] and test the average accuracy and the F1-score on the query set. The final results are reported as the mean classification accuracy over 600 randomly sampled episodes with 95 confidence intervals.

C. Results

1) Results on SD-198 Dataset: The results of our proposed SCAN approach together with several other different methods on the SD-198 dataset are reported in Tables I and II. Table I shows the accuracy and F1-score with two-way one-shot and two-way five-shot settings, and Table II lists the sensitivity, specificity, accuracy, and F1-score with five-way one-shot and five-way five-shot settings. Among the compared methods, PCN [14] and Meta-derm [16] are two meta-learning-based FSL algorithms designed for rare skin disease classification. PCN [14] uses the subcluster idea, which is similar to ours. However, it divides each class into the same number of subclusters, which fails to represent the unique structures of different classes. When we consider the diverse numbers of subclusters for various classes, the model learns more accurate feature embeddings for the base set and performs better results

TABLE III

TWO-WAY ONE-SHOT AND TWO-WAY FIVE-SHOT FEW-SHOT CLASSIFICATION RESULTS (%) ON THE DERM7PT DATASET. RESULTS OF OTHER METHODS ARE PRODUCED BY OUR REIMPLEMENTATION. †RESULTS REPORTED IN THIS ARTICLE

Method	Backbone	2-way 1-shot		2-way 5-shot	
		Accuracy	F1-score	Accuracy	F1-score
PCN [14]	Conv4	59.98±1.28	58.54±1.63	70.62±1.38	71.85±1.48
SCAN (ours)		61.42±1.49	61.90±1.66	72.58±1.28	74.05±1.32
Meta-derm [16]	Conv6	61.8†	-	76.9†	-
SCAN (ours)		62.80±1.34	63.75±1.50	76.65±1.21	73.60±1.25
NCA [42]	WRN-28-10	56.32±1.29	56.41±1.46	67.18±1.15	68.13±1.22
Baseline [39]		59.43±1.34	59.61±1.50	74.28±1.14	75.26±1.17
S2M2_R [43]		61.37±1.33	61.52±1.52	79.83±1.34	80.69±1.36
NegMargin [44]		58.00±1.44	57.50±1.65	70.12±1.30	71.07±1.36
PT+NCM [46]		60.92±1.68	61.12±1.73	74.33±1.48	74.96±1.51
PEM$_b$E-NCM [47]		60.40±1.72	60.57±1.77	72.63±1.48	73.01±1.50
EASY [48]		61.02±1.67	61.25±1.71	75.98±1.41	76.43±1.43
SCAN (ours)		**66.75±1.35**	**67.71±1.45**	**82.57±1.13**	**83.73±1.12**

TABLE IV

ABLATION STUDIES ON THE IMPROVEMENTS WITH VARIOUS BACKBONE ARCHITECTURES ON THE SD-198 DATASET. THE ACCURACY (%) OF TWO-WAY AND FIVE-WAY SETTINGS ARE DISPLAYED

Method	Backbone	2-way 1-shot	2-way 5-shot	5-way 1-shot	5-way 5-shot
Baseline [39]	Conv4	73.98±1.38	88.67±1.01	80.62±0.41	89.40±0.36
SCAN (ours)		77.12±1.44	90.22±0.95	81.85±0.41	89.79±0.34
Baseline [39]	Conv6	73.48±1.43	86.35±1.08	80.03±0.42	87.63±0.37
SCAN (ours)		76.75±1.42	87.45±1.08	81.37±0.42	89.53±0.36
Baseline [39]	ResNet18	71.52±1.45	84.22±1.15	78.95±0.41	86.87±0.37
SCAN (ours)		73.57±1.53	85.83±1.13	80.19±0.40	88.30±0.34
Baseline [39]	ResNet34	71.92±1.45	84.53±1.17	78.83±0.40	86.80±0.37
SCAN (ours)		73.08±1.42	85.03±1.12	79.54±0.42	87.45±0.40
Baseline [39]	WRN-28-10	75.72±1.47	88.95±1.00	80.63±0.42	89.46±0.38
SCAN (ours)		80.20±1.44	91.48±0.88	83.23±0.44	92.16±0.31

on the novel set. Meta-derm [16] is built on Reptile and ProtoNets by using a substitute group equivariant convolutions, and they present the state-of-the-art accuracy on the SD-198 dataset with two-way one-shot and five-shot settings. They do not consider the large intraclass variation issue for skin diseases; therefore, our method exceeds the accuracy of the Meta-derm method by 11.45% and 3.75% in two-way one-shot and two-way five-shot experiments, respectively.

Apart from the meta-learning-based methods, we also conduct comparison experiments on seven state-of-the-art transfer-learning-based FSL methods. NCA [42] proposes neighbor component analysis loss, which can learn the relationships between features. The Baseline [39] method was proposed as the baseline model for transfer-learning-based FSL methods. It trains the feature encoder with a standard cross-entropy loss. S2M2_R [43] utilizes self-supervision and Manifold Mixup to enhance the feature encoder. NegMargin [44] applies a negative margin softmax loss when training on the base set and obtains the increased results on the novel set. Instead of training a powerful feature encoder in the pretraining step, some methods apply feature postprocessing tricks in the fine-tuning step. PT+NCM [46] proposes power transform (PT) on the support and the query features to align their distributions closer to Gaussian-like distributions. In addition to PT, PEM$_b$E [47] applies the Euclidean normalization and mean subtraction to further reduce the task bias. EASY [48] uses random resized crops to augment support set images and leverages the ensemble tricks to further boost the performance. Compared to them, SCAN pays more attention to the feature encoder training instead of the postprocessing tricks. We propose the specific unsupervised cluster branch to handle the inherent subcluster issue in skin disease datasets. Therefore, our model learns more precise feature embeddings on the base set and shows better performance on the novel set.

2) Results on Derm7pt Dataset: The results of our proposed SCAN method and other compared methods on the Derm7pt dataset are presented in Table III. We compare SCAN with two state-of-the-art meta-learning and seven transfer-learning FSL methods. In general, the experimental results show that our method provides superior performance on the Derm7pt dataset among all the compared algorithms. For the setting of two-way five-shot, SCAN exceeds all the compared methods except for Meta-derm [16] when using Conv6 as the backbone architecture.

D. Ablation Study

1) Backbone Architectures: In addition to the WRN-28-10 architecture, we conducted performance tests of the SCAN method on Conv4, Conv6, ResNet18, and ResNet34 architectures to assess its robustness across different backbone networks. We compared SCAN with the Baseline method [39], which utilizes only the $\mathcal{L}_{\text{class}}$ loss. The results are presented in Table IV. Our proposed method, SCAN, consistently improves the performance across various backbone networks. However,

TABLE V

ABLATION STUDIES ON THE EFFECTIVENESS OF CLUSTER BRANCH AND THE PURITY LOSS ON THE SD-198 DATASET. THE ACCURACY (%) OF TWO- AND FIVE-WAY SETTINGS AND THE CLUSTER ERROR RATE (%) ARE REPORTED

Class branch ($\mathcal{L}_{class}$)	Cluster branch ($+\mathcal{L}_{cluster}$)	Cluster branch ($+\mathcal{L}_{purity}$)	2-way 1-shot	2-way 5-shot	5-way 1-shot	5-way 5-shot	Cluster Error Rate
✓			75.72±1.47	88.95±1.00	80.63±0.42	89.46±0.38	-
	✓		63.37±1.46	75.28±1.25	71.58±0.42	78.54±0.39	61.03
✓	✓		78.65±1.46	89.46±0.98	81.29±0.42	90.88±0.37	35.10
✓	✓	✓	**80.20±1.44**	**91.48±0.88**	**83.23±0.44**	**92.16±0.31**	**20.58**

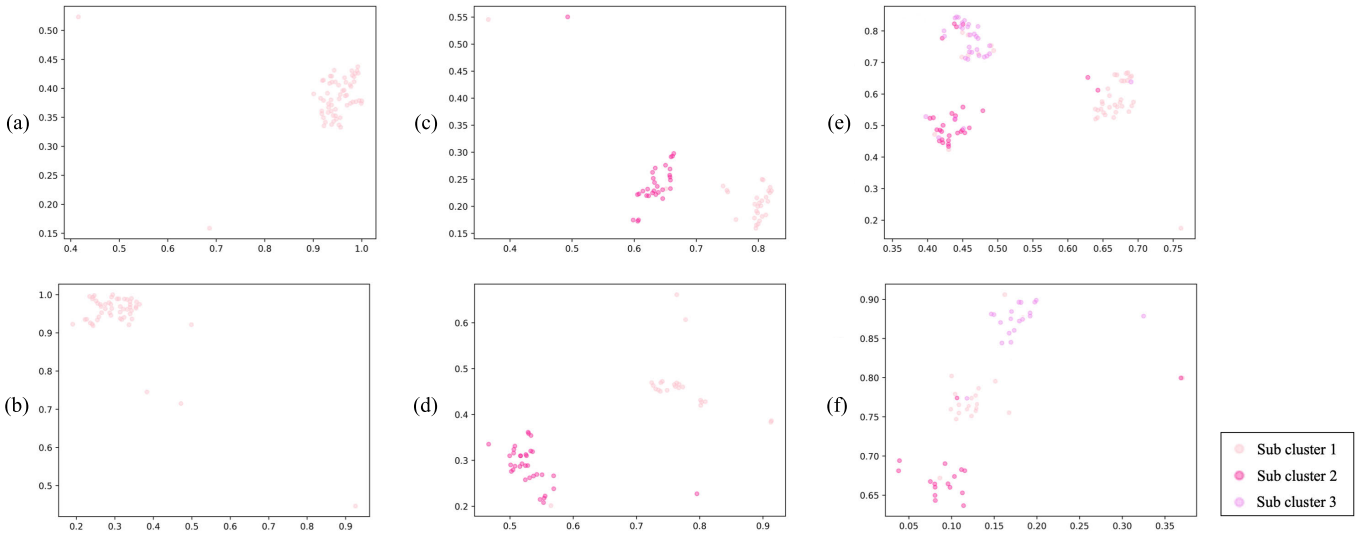

Fig. 6. t-SNE visualization results of feature embeddings learned by SCAN. Six classes in the base set of the SD-198 dataset: (a) perioral dermatitis, (b) rhinophyma, (c) psoriasis, (d) sebaceous gland hyperplasia, (e) actinic solar damage (actinic keratosis), and (f) epidermoid cyst. Various numbers of subclusters are identified for different classes.

we observed that the performance of ResNet backbones generally falls behind that of Conv backbones. This discrepancy may be attributed to overfitting on the base set. By increasing the width of the residual blocks, the overfitting problem is mitigated. The best performance is achieved when employing the WRN backbone.

2) Effectiveness of Each Component: We conducted evaluations to assess the effectiveness of each component in the SCAN framework. The results are presented in Table V and include four scenarios: 1) using only the class branch; 2) using only the cluster branch; 3) using both the class and cluster branches without the purity loss; and 4) using both the class and cluster branches with the purity loss. In addition to reporting the accuracy of the novel set, we introduced a metric called the cluster error rate to evaluate the quality of the learned subclusters on the base set. The cluster error rate measures the percentage of samples that are assigned to incorrect clusters out of the entire dataset. It is calculated by summing the number of items with different class labels from the major examples for each cluster and dividing it by the total size of the training set.

Based on the results in Table V, we highlight the following three key findings. First, models trained with only the class branch or the cluster branch exhibit poor performance, with two-way one-shot accuracy of 75.72% and 63.37%, respectively. The cluster branch performs worse than the class branch, likely due to the absence of supervised class labels in the cluster branch. Second, combining both the class and cluster branches leads to a significant improvement in accuracy. For instance, the accuracy of the two-way one-shot setting is enhanced from 75.72% to 78.65%. This indicates the effectiveness of addressing the challenge posed by large intraclass variation. Third, the inclusion of the purity loss helps refine the clustering results, resulting in a significant decrease in the cluster error rate. This demonstrates the ability of the purity loss to enhance the quality of the learned subclusters. Overall, these findings confirm the effectiveness of the different components in the SCAN framework for learning dynamic subcluster structures and improving the classification performance.

3) Visualization of Feature Distribution in Base Classes: To showcase the capability of our proposed SCAN framework in learning the dynamic subcluster structures, we utilize t-SNE visualization [57] to depict the feature distributions within the base classes of the SD-198 dataset. Fig. 6 presents the t-SNE visualization of six base classes, where the pseudo subcluster labels learned by SCAN are leveraged to illustrate the feature distribution within each class. The subcluster structures shown in Fig. 6 reveal varying numbers of subclusters for each class, ranging from one to three. This dynamic identification of subcluster numbers for different classes demonstrates the capability of SCAN.

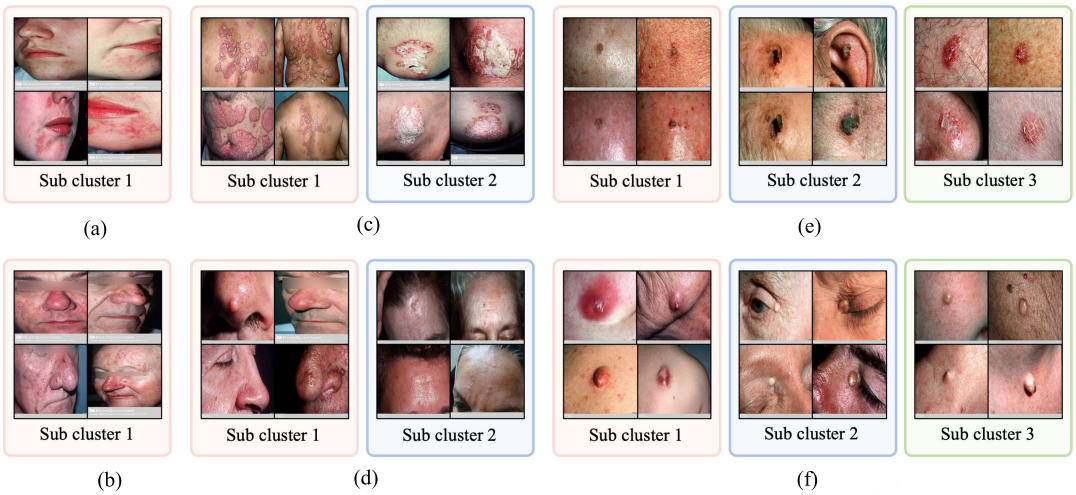

Fig. 7. Examples of subclusters learned by SCAN. Images from six classes in the base set of the SD-198 dataset: (a) perioral dermatitis, (b) rhinophyma, (c) psoriasis, (d) sebaceous gland hyperplasia, (e) actinic solar damage (actinic keratosis), and (f) epidermoid cyst.

TABLE VI

DISCRIMINATIVE ANALYSIS ON THE BASE AND NOVEL CLASSES OF SD-198 DATASET. THE INTERCLASS DISTANCE (D_{inter}), THE INTRACLASS DISTANCE (D_{intra}), AND THE DISCRIMINATIVE INDEX (ϕ) ARE COMPUTED

Class branch	Cluster branch	Base classes			Novel classes		
		D_{inter}	D_{intra}	ϕ	D_{inter}	D_{intra}	ϕ
✓		1.7251	0.1801	9.5777	0.7631	0.6235	1.2283
	✓	0.5990	0.6964	0.8601	0.7807	0.6739	1.1585
✓	✓	1.6377	0.5067	3.2321	0.8318	0.5893	1.4115

To validate the rationality of the clustering outcomes, we display some representative images from each subcluster for the six base classes in Fig. 7(a)–(f). For the classes perioral dermatitis and rhinophyma, the entire class is considered a unified group due to the high similarity among items. As shown in Fig. 7(a) and (b), the affected regions of these two classes are mostly concentrated in the same area (the chin region for perioral dermatitis and the nose area for rhinophyma). In cases where a class forms multiple subclusters, the images within different subclusters originate from distinct body locations, resulting in diverse symptoms.

In summary, our proposed method effectively learns the cluster arrangement within each class through the incorporation of an additional unsupervised cluster branch and two restricted losses. This enables the dynamic exploration of subcluster structures within the dataset.

4) Discriminability Analysis on Feature Embeddings: In this section, we explore the impact of the proposed method on the features in the base and novel set. In the pretraining stage on the base classes, the class branch aims to increase the distance among different class centers (i.e., the interclass distance), while the cluster branch aims to increase the distances within a class (i.e., the intraclass distance).

To quantitatively measure the impact of two branches on both the base and novel classes, we compute the interclass distance (D_{inter}), the intraclass distance (D_{intra}), and the discriminative index (ϕ), based on the definitions provided in [44] and [58]. The equations of three metrics are

$$D_{\text{inter}} = \frac{1}{k(k-1)} \sum_{m=1}^{k} \sum_{n=1, n\neq m}^{k} \|\mu(c_m) - \mu(c_n)\|_2^2 \tag{8}$$

$$D_{\text{intra}} = \frac{1}{k} \sum_{m=1}^{k} \left(\frac{1}{|c_m|} \sum_{(x_i, y_i)\in c_m}^{k} \left\| \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|} - \mu(c_m) \right\|_2^2 \right) \tag{9}$$

$$\phi = \frac{D_{\text{inter}}}{D_{\text{intra}}} \tag{10}$$

where $\mu(c_m)$ denotes the center of class c_m and is calculated by the mean of the $L2$-normalized feature embeddings of class c_m.

D_{inter} is computed by the mean $L2$ distances between every pair of class centers. D_{intra} is computed by the mean $L2$ distances between every sample in a class and its corresponding class center. The discriminative index ϕ is defined as the division of D_{inter} and D_{intra}. The higher value of ϕ indicates more discriminative feature embeddings, as it indicates a larger ratio of interclass distance and intraclass distance.

We present the results of D_{inter}, D_{intra}, and ϕ on the base and novel classes of the SD-198 dataset in Table VI. For the base classes, we observe that the class branch increases the interclass distance from 0.5990 to 1.6377, and the cluster branch increases the intraclass distance from 0.1801 to 0.5067. These results validate the roles of two branches on the base classes, as mentioned in our previous intuitive explanation. As a result of the increased intraclass distance and relatively unchanged interclass distance, the discriminative index decreases for the base classes when adding the cluster branch to the class branch. However, we argue that less discrimination in base set features does not indicate a conflict between two branches, as our primary focus is the performance of the novel classes. For the novel classes, the combination of two branches leads to an increase in interclass distance and a decrease in

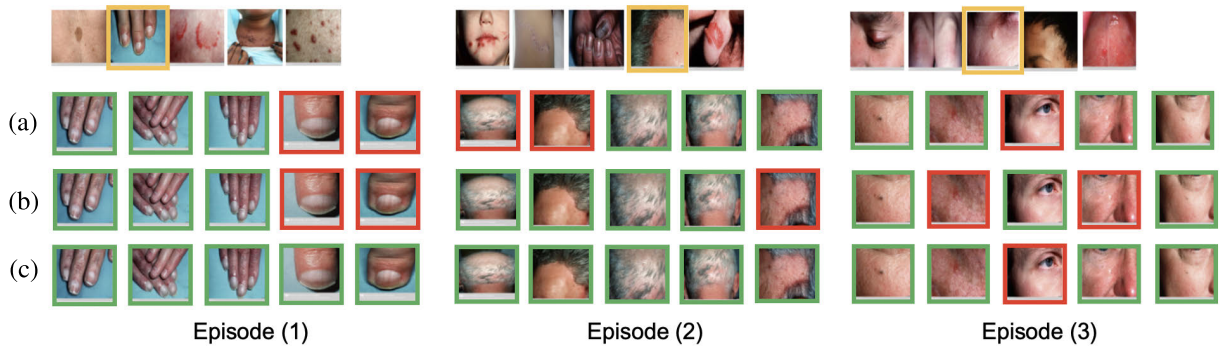

Fig. 8. Results of query images prediction on the SD-198 novel set for the five-way one-shot task. Episodes (1)–(3) are sampled episodes. For each episode, in the top row are images from the support set. There are five classes in total and one image from each class. The remaining five images are the query images from one of the above five classes in the support set, and we use a yellow bounding box to show the target class. Three rows show the testing results of three methods. (a) Baseline [39]. (b) PCN [14]. (c) SCAN (ours). The green/red bounding boxes denote the true/false classification.

intraclass distance compared to using either the class or the cluster branch alone.

These results demonstrate that our proposed method enables the feature encoder to embed more generalizable feature representations by learning the subcluster structures of the base classes. Consequently, the feature encoder can extract more discriminative features for the novel classes, resulting in the improved classification accuracy.

5) Visualization of Classification Results on Novel Classes: To compare the episodic predictions among various methods, we display three examples for the five-way, one-shot experiment of Baseline [39], PCN [14], and our proposed SCAN method, as shown in Fig. 8. The first row shows the sampled support set images, with one example per class from a total of five classes. The yellow bounding box indicates the ground truth category for the query images. The subsequent three rows illustrate the results of the query images for each method. The green and red boxes represent correct and incorrect predictions, respectively, on the target class in the support set.

In general, our method demonstrates higher accuracy in predicting the listed query set compared to the other methods, particularly evident in episodes (1) and (2) of Fig. 8. While our method correctly identifies examples that exhibit high similarity to the support image, such as the first three images in (1), it also effectively handles cases that are distinguishable from the given support item. For instance, it accurately predicts the last two columns in (1) and the first and fifth columns in (2). However, it is worth noting that the incorrect prediction in the third column of episode (3) suggests that the performance may be influenced by large interclass similarity. This is evident as the third query image bears more resemblance to the support image from the first class than to the image sampled from its own class.

6) Ablation Studies on Hyperparameters: We conducted additional experiments to assess the efficacy of the hyperparameters in SCAN. Fig. 9 presents the results of two-way one-shot and two-way five-shot on two datasets, demonstrating their correlation with the cluster number. Our findings indicate that our method enhances the overall performance across a range of cluster numbers. Notably, cluster numbers that are either too small or too large fail to adequately represent the subcluster structures. The highest accuracy is attained when

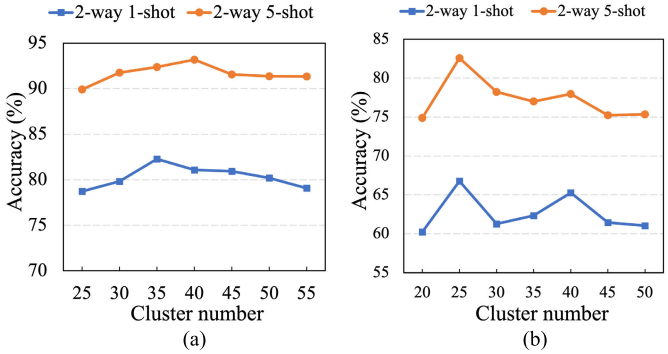

Fig. 9. Ablation studies on the hyperparameter in SCAN. The accuracy of two-way one-shot and two-way five-shot is plotted with respect to the cluster number for (a) SD-198 and (b) Derm7pt datasets.

TABLE VII
COMPUTATIONAL EFFICIENCY ANALYSIS OF THE PROPOSED METHOD

Method	Time per epoch	FLOPs	Model parameters
Baseline [39]	12.28 sec	32.81 GMac	36.60 M
PCN [14]	16.44 sec	32.81 GMac	36.60 M
SCAN (ours)	20.93 sec	32.81 GMac	37.12 M

the cluster number is 40 for the SD-198 dataset and 25 for the Derm7pt dataset.

7) Computational Efficiency Analysis: To evaluate the computational efficiency of our proposed method, we conducted a comparison with the Baseline [39] and PCN [14] methods on the SD-198-20 dataset. The Baseline and PCN methods have identical network architectures, which only contain the class branch. We utilized the same backbone, hyperparameters, and GPU for all methods. The results of time, floating-point operations (FLOPs), and model parameters are reported in Table VII. We trained all three methods for 800 epochs and recorded the average time consumed for one epoch. Our proposed SCAN method takes approximately 8 s longer than the Baseline method and 4 s longer than the PCN method to complete the clustering calculations. However, our performance is vastly superior to the Baseline and PCN methods, improving the accuracy by 4.48% and 7.09%, respectively. We also computed the number of FLOPs for all methods, using an input size of $3 \times 80 \times 80$ and a batch size of 1.

The results show that the FLOPs numbers are nearly the same for all three methods. The additional computational effort of the linear classifier in the cluster branch of SCAN can be considered negligible. Finally, we compared the number of parameters of SCAN with the other two methods. The SCAN method has approximately 0.52 M more parameters than the Baseline and PCN methods. These additional parameters in SCAN are derived from the extra linear classifier in the cluster branch.

Overall, our proposed method achieves superior performance compared to the Baseline and PCN methods, while the additional computational cost is minimal. This highlights the effectiveness and efficiency of our proposed method for rare skin disease diagnosis.

V. Conclusion

In this article, we present a novel SCAN for rare skin disease classification. According to our key insights, skin disease datasets have existing latent subgroups within a class. Therefore, our proposed SCAN effectively learns the intrinsic subcluster structures of skin disease via a well-designed dual-branch framework and three additive losses. The results tested on two public skin image datasets show that our method excels over other state-of-the-art methods by around 2%–5% for various settings.

References

[1] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*, 2019, pp. 605–613.

[2] X. Li et al., "3Dmulti-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images," *Med. Image Anal.*, vol. 45, pp. 41–54, Apr. 2018, pp. 424–432.

[3] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2016.

[4] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[5] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[6] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[7] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[8] J. Sun, S. Lapuschkin, W. Samek, Y. Zhao, N.-M. Cheung, and A. Binder, "Explanation-guided training for cross-domain few-shot classification," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7609–7616.

[9] R. Feng, X. Liu, J. Chen, D. Z. Chen, H. Gao, and J. Wu, "A deep learning approach for colonoscopy pathology WSI analysis: Accurate segmentation and classification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3700–3708, Oct. 2021.

[10] J. Chen et al., "A transfer learning based super-resolution microscopy for biopsy slice images: The joint methods perspective," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 103–113, Jan. 2021.

[11] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, and P.-A. Heng, "CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1483–1493, May 2020.

[12] B. H. Y. Chung, J. F. T. Chau, and G. K.-S. Wong, "Rare versus common diseases: A false dichotomy in precision medicine," *NPJ Genomic Med.*, vol. 6, no. 1, pp. 1–5, 2021.

[13] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 206–222.

[14] V. Prabhu, A. Kannan, M. Ravuri, M. Chaplain, D. Sontag, and X. Amatriain, "Few-shot learning for dermatological disease diagnosis," in *Proc. Mach. Learn. Healthcare Conf. (MLHC)*, 2019, pp. 532–552.

[15] X. Li, L. Yu, Y. Jin, C.-W. Fu, L. Xing, and P.-A. Heng, "Difficulty-aware meta-learning for rare disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2020, pp. 357–366.

[16] K. Mahajan, M. Sharma, and L. Vig, "Meta-DermDiagnosis: Few-shot skin disease identification using meta-learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3142–3151.

[17] D. Zhang, M. Jin, and P. Cao, "ST-MetaDiagnosis: Meta learning with spatial transform for rare skin disease diagnosis," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 2153–2160.

[18] W. Zhu, H. Liao, W. Li, W. Li, and J. Luo, "Alleviating the incompatibility between cross entropy loss and episode training for few-shot skin disease classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2020, pp. 330–339.

[19] W. Zhu, W. Li, H. Liao, and J. Luo, "Temperature network for few-shot learning with distribution-aware large-margin metric," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107797.

[20] R. Singh et al., "MetaMed: Few-shot medical image classification using gradient-based meta-learning," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108111.

[21] C. Zhou, M. Sun, L. Chen, A. Cai, and J. Fang, "Few-shot learning framework based on adaptive subspace for skin disease classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2022, pp. 2231–2237.

[22] Y. Guo et al., "A broader study of cross-domain few-shot learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 124–141.

[23] D. Chen, Y. Chen, Y. Li, F. Mao, Y. He, and H. Xue, "Self-supervised learning for few-shot image classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1745–1749.

[24] C. Medina, A. Devos, and M. Grossglauser, "Self-supervised prototypical transfer learning for few-shot classification," 2020, *arXiv:2006.11325*.

[25] C. Perng Phoo and B. Hariharan, "Self-training for few-shot transfer across extreme task differences," 2020, *arXiv:2010.07734*.

[26] Z. Dai et al., "PFEMed: Few-shot medical image classification using prior guided feature enhancement," *Pattern Recognit.*, vol. 134, Feb. 2023, Art. no. 109108.

[27] J. Xiao, H. Xu, D. Fang, C. Cheng, and H. Gao, "Boosting and rectifying few-shot learning prototype network for skin lesion classification based on the Internet of medical things," *Wireless Netw.*, vol. 29, no. 4, pp. 1507–1521, May 2023.

[28] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1126–1135.

[29] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," 2017, *arXiv:1703.05175*.

[30] V. Anand, S. Gupta, and D. Koundal, "Skin disease diagnosis: challenges and opportunities," in *Proc. 2nd Doctoral Symp. Comput. Intell.*, 2022, pp. 449–459.

[31] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 538–546, Mar. 2019.

[32] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv:1803.02999*.

[33] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3630–3638.

[34] X. Li, L. Yu, C.-W. Fu, M. Fang, and P.-A. Heng, "Revisiting metric learning for few-shot image classification," *Neurocomputing*, vol. 406, pp. 49–58, Sep. 2020.

[35] Y. Liu, L. Zhu, X. Wang, M. Yamada, and Y. Yang, "Bilaterally normalized scale-consistent sinkhorn distance for few-shot image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 17, 2023, doi: 10.1109/TNNLS.2023.3262351.

[36] N. Lai, M. Kan, C. Han, X. Song, and S. Shan, "Learning to learn adaptive classifier–predictor for few-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3458–3470, Aug. 2021.

[37] S. An, S. Kim, P. Chikontwe, and S. H. Park, "Dual attention relation network with fine-tuning for few-shot EEG motor imagery classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 28, 2023, doi: 10.1109/TNNLS.2023.3287181.

[38] Y. Zhang, S. Huang, X. Peng, and D. Yang, "Semi-identical twins variational AutoEncoder for few-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 9, 2023, doi: 10.1109/TNNLS.2022.3233553.

[39] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Frank Wang, and J.-B. Huang, "A closer look at few-shot classification," 2019, *arXiv:1904.04232*.

[40] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," 2019, *arXiv:1909.02729*.

[41] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 266–282.

[42] Z. Wu, A. A. Efros, and S. X. Yu, "Improving generalization via scalable neighborhood component analysis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 685–701.

[43] P. Mangla, M. Singh, A. Sinha, N. Kumari, V. N. Balasubramanian, and B. Krishnamurthy, "Charting the right manifold: Manifold mixup for few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2207–2216.

[44] B. Liu et al., "Negative margin matters: Understanding margin in few-shot classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 438–455.

[45] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," 2021, *arXiv:2101.06395*.

[46] Y. Hu, V. Gripon, and S. Pateux, "Leveraging the feature distribution in transfer-based few-shot learning," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, 2021, pp. 487–499.

[47] Y. Hu, S. Pateux, and V. Gripon, "Squeezing backbone feature distributions to the max for efficient few-shot learning," *Algorithms*, vol. 15, no. 5, p. 147, Apr. 2022.

[48] Y. Bendou et al., "Easy—Ensemble augmented-shoty-shaped learning: State-of-the-art few-shot classification with simple components," *J. Imag.*, vol. 8, no. 7, p. 179, 2022.

[49] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, Jan. 2023.

[50] L. Zhang et al., "TN-ZSTAD: Transferable network for zero-shot temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3848–3861, Mar. 2023.

[51] M. Li, P.-Y. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2023.

[52] C. Yan et al., "ZeroNAS: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9733–9740, Dec. 2022.

[53] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "SimpleShot: Revisiting nearest-neighbor classification for few-shot learning," 2019, *arXiv:1911.04623*.

[54] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 2990–2999.

[55] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[58] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Proc. IEEE Signal Process. Soc. Workshop*, 1999, pp. 41–48.

Shuhan Li received the bachelor's degree from the City University of Hong Kong, Hong Kong, China, in 2019. She is currently pursuing the Ph.D. degree with the Department of Computer Science Engineering, Hong Kong University of Science and Technology, Hong Kong.

Her research interests include artificial intelligence and medical image analysis.

Xiaomeng Li (Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong, Shenzhen, China, in 2019.

She is an Assistant Professor with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong. Prior to joining HKUST, she worked as a Post-Doctoral Researcher at Stanford University, Stanford, CA, USA. Her research primarily revolves around artificial intelligence and medical image analysis, with a specific emphasis on leveraging machine intelligence to advance healthcare.

Xiaowei Xu received the B.S. and Ph.D. degrees in electronic science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2011 and 2016 respectively.

He is with the Department of Cardiovascular Surgery, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Southern Medical University. His research interests include deep learning, and medical image segmentation.

Dr. Xu was a recipient of DAC system design contest special service recognition reward in 2018.

Kwang-Ting Cheng (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, CA, USA, in 1988.

He holds the position of Vice-President for Research and Development at the Hong Kong University of Science and Technology (HKUST). Additionally, he has successfully transferred several of his inventions into commercial products. His contributions have garnered recognition from the academic community. He has an extensive publication record, with over 500 technical papers, co-authorship of five books, and 12 US patents. His research interests primarily lie in computer vision, medical image analysis, and electronic design automation

Dr. Cheng is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) and the Hong Kong Academy of Engineering Sciences (HKAES).