

# Hyperbolic Geometry-Driven Robustness Enhancement for Rare Skin Disease Diagnosis

Yang Hu, Yuanyuan Chen, Xiaohan Xing, Jingfeng Zhang, Bolysbek Murat Yerzhanuly<sup>✉</sup>,  
Bazargul Matkerim, and Yong Xia<sup>✉</sup>, *Member, IEEE*

**Abstract**—The automated diagnosis of rare skin diseases using dermoscopy images, known as a few-shot learning (FSL) problem, remains challenging, since traditional FSL research tends to disregard the intrinsic hierarchical nature of rare diseases and data uncertainty. To address these issues, we propose to conduct rare skin disease diagnosis in hyperbolic space, which facilitates implicit class hierarchical structures and precise uncertainty measurement due to pivotal geometrical properties. We propose a *Hyperbolic Geometry-driven Robustness Enhancement (HGRE)* framework specifically tailored for diagnosing rare skin diseases. The HGRE framework uses implicit hierarchical relation in the hyperbolic space to better represent the features of rare diseases. Moreover, the framework incorporates an *Adversarial Proxy Construction (APC)* module to address the problem of data uncertainty. Specifically, the APC module uses the distance to the hyperbolic space origin as an indicator of uncertainty to filter and construct adversarial proxies for each uncertain prototype to achieve adversarial robust training. Leveraging the two unique geometrical properties, our HGRE framework effectively addresses the limitations of insufficient hierarchical relation utilization and data uncertainty in FSL-based rare skin disease diagnosis. This enhancement of the model's robustness in training has been corroborated by extensive empirical validation

on two skin lesion datasets, where HGRE's performance notably surpassed existing state-of-the-art FSL methods.

**Index Terms**—Few-shot classification, rare skin diseases, hyperbolic space, adversarial learning.

## I. INTRODUCTION

WITH the advancements in deep learning techniques, we have witnessed significant progress in developing computer-aided diagnostic tools for skin diseases [1], [2], [3], [4]. Nevertheless, while deep learning has exhibited remarkable performance in the diagnosis of common skin diseases, its efficacy remains constrained when it comes to rare skin diseases owing to the inadequacy of training samples. Due to the low incidence rates and complex patterns associated with rare skin diseases, physicians face considerable challenges in diagnosing rare skin diseases. Therefore, creating an accurate diagnostic tool for rare skin diseases holds greater clinical value than models for common diseases.

To improve the diagnostic accuracy of rare diseases, existing methods have explored training deep models with limited data using few-shot learning (FSL) techniques [5], [6], [7], [8], [9], [10]. In this context, rare diseases with extremely limited training samples are treated as novel classes, while common diseases with abundant training data serve as base classes. FSL-based methods [5], [6], [7], [8], [9], [10] address the diagnosis of rare diseases by acquiring knowledge from abundant data on common diseases and transferring it to the rare disease domain. Augmentation-based FSL methods [11], [12], [13], [14] tackle the problem from a data augmentation perspective. These methods augment the features of few-shot samples by transferring either the intra-class variance [13], [14] or feature distribution [11], [12] from base classes to novel classes. While effective in augmenting data for rare diseases, these methods necessitate predefined rare disease classes during training and require substantial time for feature augmentation and model fine-tuning, limiting their clinical deployment flexibility.

Metric-learning-based approaches [15], [16], [17], [18] represent another class of methods capable of addressing the diagnosis of rare diseases. They simulate the few-shot scenario by iteratively sampling a small dataset from base classes data. After training with base classes data, these methods can effectively classify novel classes with only a small number of supporting samples without extensive retraining, which greatly enhances the flexibility of clinical deployment. Among them,

Received 26 March 2024; revised 17 October 2024; accepted 13 November 2024. Date of publication 21 November 2024; date of current version 7 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62171377, Grant 62401452 and Grant 92470101, in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20241360, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LTGY23H180003, in part by the Key Project of Ningbo Public Welfare Science and Technology under Grant 2021S107, in part by the Ningbo Clinical Research Center for Medical Imaging through Open Project 2022LYKFZD06 under Grant 2021L003, and in part by the Fundamental Research Funds for the Central Universities. (Yang Hu and Yuanyuan Chen contributed equally to this work.) (Corresponding author: Yong Xia.)

Yang Hu and Yong Xia are with the School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yhu@mail.nwpu.edu.cn; yxia@nwpu.edu.cn).

Yuanyuan Chen is with the Department of Cell Biology and Genetics, School of Basic Medical Sciences, Xi'an Jiaotong University Health Science Center, Xi'an 710061, China (e-mail: yychen@mail.xjtu.edu.cn).

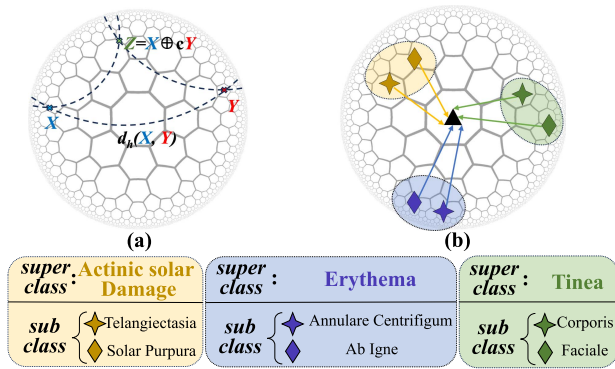
Xiaohan Xing is with the Department of Radiation Oncology, Stanford University, Stanford, CA 94305 USA (e-mail: xhxing@stanford.edu).

Jingfeng Zhang is with Ningbo No.2 Hospital, Ningbo 315000, China (e-mail: jingfengzhang73@zjhu.edu.cn).

Bolysbek Murat Yerzhanuly is with Northwestern Polytechnical University Kazakhstan Branch, Almaty 050040, Kazakhstan (e-mail: muratbolysbek@mail.nwpu.edu.cn).

Bazargul Matkerim is with Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan (e-mail: Bazargulmm@gmail.com).

Digital Object Identifier 10.1109/JBHI.2024.3500094



**Fig. 1.** Introduction to the principles of the Poincaré ball. (a) Dotted lines represent geodesics, and the length of the geodesic represents the distance between two points in the hyperbolic space, such as  $d_h(X, Y)$ . Point  $Z$  represents the Möbius addition of points  $X$  and  $Y$ . (b) In the hyperbolic space, subclasses belonging to a superclass (represented by the same color) converge in similar directions, while the directions of different superclasses vary significantly.

ProtoNet [15], as a representative classical algorithm, is often utilized as a baseline model for various methods [16], [19], [20]. It learns a metric space where the prototype representation of each class is generated by averaging the features of the support samples. Subsequently, classification is executed by evaluating the distance from the query samples to each prototype.

Despite their promise in diagnosing rare skin diseases, metric-learning-based methods encounter challenges in two critical areas affecting classification accuracy: **distance metrics** and **prototype calculations**. First, skin disease datasets often exhibit hierarchical structures in the label space, providing valuable prior knowledge about class relationships. Capturing this hierarchical structure is essential for distance metrics, especially when training data is scarce. However, most existing methods measure distance metrics based on the Euclidean space, which struggles to effectively represent this hierarchy without explicit constraints [8], [9], [10]. Second, due to limited data and significant intra-class variance in skin images, some noisy or rarely observed samples exhibit features that deviate markedly from other samples. These atypical samples may yield lower prediction confidence, introducing uncertainty in determining the prototype for that class. Understanding and modeling prototype uncertainty can enhance FSL method optimization, thereby improving model robustness.

To address these challenges, we suggest conducting rare skin disease diagnosis in **hyperbolic space**, which facilitates distance metrics and prototype calculations due to the following geometrical properties. First, the hyperbolic space (as shown in Fig. 1(a)) outperforms Euclidean space in representing hierarchical relationships, particularly in tree-like data structures [21], [22], [23]. By harnessing this potent representational capability, computing distance metrics in hyperbolic space enhances the expression of similarity between samples and classes. Subclasses under the same superclass are relatively close to each other while maintaining a certain distance, whereas the distance between different superclasses remains significant (see Fig. 1(b)). Second, in the hyperbolic space, there exists a learning tendency: more

generic or ambiguous embeddings are closer to the center, while more specific embeddings move toward the boundary during training [21], [23]. Consequently, the distance to the center provides a refined estimation of feature uncertainty, enabling measurement of prototype uncertainty and uncertainty-aware FSL optimization. Benefiting from these unique geometrical properties of hyperbolic space, we simultaneously optimize distance metrics and prototype calculations. This method promotes a more robust FSL algorithm for rare skin disease diagnosis, leading to improved accuracy and robustness.

In this paper, we propose the *Hyperbolic Geometry-driven Robustness Enhancement (HGRE)* framework for rare skin disease diagnosis. While adhering to the conventional metric-based FSL paradigm, our HGRE framework diverges in the choice of metric space. Instead of relying on Euclidean space [9], [15], HGRE maps image features from the Euclidean space into hyperbolic space and conducts metric learning within this domain. By doing so, the computed distance metrics implicitly consider the hierarchical relationships between classes, significantly enhancing the model's ability to handle rare skin diseases. To further address the data uncertainty inherent in rare skin diseases, we introduce an Adversarial Prototype Correction (APC) module within the HGRE framework. Initially, this module filters uncertain prototypes by calculating their distance from the hyperbolic origin. Subsequently, it generates adversarial proxies for each uncertain prototype and conducts adversarial training based on these proxies. This method effectively mitigates the impact of uncertain data on the model's robustness.

The main contributions of our work are threefold:

1. We utilize hyperbolic space's ability to extract hierarchical representations, modeling the relationships between different classes of rare skin diseases, which enhances model performance by exploring these inter-class relationships.
2. We leverage the property in hyperbolic space where the distance from features to the origin serves as an estimate of uncertainty, enabling uncertainty-based prototype selection and adversarial prototype construction.
3. We demonstrate the effectiveness of the proposed HGRE framework on the SD-198 dataset and the ISIC 2019 skin lesion datasets, showcasing superior performance compared to state-of-the-art methods.

## II. RELATED WORK

### A. Few-Shot Image Classification

Few-shot image classification has garnered significant research attention, leading to the development of various methods. These methods can be broadly categorized into two main groups: **data augmentation techniques** and **meta-learning techniques**.

Data augmentation techniques [11], [12], [13], [14] aim to expand the training data by generating synthetic samples, modeling data distribution, or applying other augmentation strategies. By augmenting the training data, these methods seek to enhance the model's performance. For example, Zhang et al. [11] modeled similarity as a Gaussian distribution and employed graph optimization combined with Monte Carlo sampling for data

augmentation. Xu et al. [12] utilized a conditional variational autoencoder model to generate visual samples based on semantic embeddings. Cao et al. [13] associated each novel class with base classes using semantic similarity, facilitating the construction of a discriminative feature space for each novel class.

Meta-learning techniques [15], [16], [17], [18], [24], [25], [26], [27] simulate the few-shot setting by iteratively sampling a small dataset from base classes. Importantly, these methods do not require novel class data during training. The model can rapidly classify novel classes with only a few supporting samples. Meta-learning has been approached from two perspectives within deep learning: (1) **Optimization-based methods:** These methods use base classes to train an initial model and then adapt the model to novel classes through a few gradient updates on a small number of labeled samples [24], [25], [26], [27]. (2) **Metric-learning-based methods:** These methods aim to construct a generalized representation space, where query are classified based on the distance of each class prototype. Various metrics have been explored in existing work, including Euclidean distance [15], cosine similarity [28], task adaptive metrics [29], metrics based on local descriptors [30], and graph neural networks [31].

Several existing methods formulate the diagnosis of rare diseases as an FSL problem [5], [6], [7], [8], [9], [10]. Among these methods, Sun et al. [10] first trained a feature extractor using abundant common disease data via unsupervised learning, and then fine-tuned this model on novel classes specifically for rare disease diagnosis. Li et al. [8] proposed a difficulty-aware meta-learning method that trains a meta-classifier on common disease data while dynamically monitoring the importance of learning tasks. Zhu et al. [9] developed a temperature network that results in a compact intra-class distribution and generates query-specific prototypes for a distribution-aware metric. Given the clinical context and the challenges associated with obtaining rare disease data, metric-learning-based FSL methods are particularly suitable for rare disease diagnosis. Most FSL methods for rare diseases rely on metric learning. However, a significant limitation lies in the fact that most existing metric learning methods operate in Euclidean space, which cannot fully leverage the implicit hierarchical relationships inherent in rare skin diseases. To address this limitation, we propose conducting rare skin disease diagnosis in the hyperbolic space. By doing so, we aim to achieve a more robust metric that better captures the unique characteristics of rare skin diseases.

## B. Hyperbolic Space

The hyperbolic space, characterized by an exponential expansion of system states, bears resemblance to hierarchical tree structures. In contrast to Euclidean space, hyperbolic space exhibits greater adaptability to complex data domains with tree-like structures such as biology, network science, and computer vision. Recent years have witnessed remarkable success in hyperbolic embeddings, particularly within the field of natural language processing [32], [33]. The intrinsic properties of hyperbolic space facilitate low-distortion embeddings of intricate structures, including tree diagrams [34], [35]. Moreover,

research has underscored its potential benefits across various computer vision problems, including the optimization of metric learning [21], [36], [37], [38].

In the realm of deep learning operating within hyperbolic space, a discernible learning trend emerges. Specifically, embeddings that exhibit greater generality or ambiguity tend to gravitate toward the center, while more narrowly defined embeddings progressively approach the periphery during training [21], [23]. Consequently, the distance to the center offers a nuanced assessment of feature uncertainty, enabling precise measurements of prototype uncertainty and the implementation of uncertainty-aware optimization in FSL. For instance, Chen et al. [36] dynamically identified pixels with misclassification and high uncertainty in hyperbolic space through hyperbolic distance calculations. Montanaro et al. [37] integrated the features of the point cloud classifier into hyperbolic space, explicitly regularizing the representation to account for partial-global levels. Khrulkov et al. [21] and Ma et al. [38] introduced hyperbolic space into FSL, conducting metric learning within this context and observing performance improvements. Anticipating the implicit hierarchical relationships inherent in hyperbolic space, we envision more effective implementations of FSL methods based on metric learning. This method holds particular promise for rare skin disease classification characterized by hierarchical structures. Moreover, hyperbolic space serves as a benchmark for uncertainty estimation in adversarial learning, facilitating the precise and elegant construction of adversarial proxies to enhance model robustness.

## C. Adversarial Learning

Adversarial learning research has long focused on the neural network classifier's vulnerability to input disturbances. The goal is to enhance standard generalization and robustness by training neural networks with adversarial examples' samples intentionally perturbed to cause erroneous model outputs [39], [40], [41], [42], [43], [44], [45]. Adversarial training involves combining these perturbed samples with original ones during training, enabling the model to better adapt to diverse input conditions. Bin et al. [39] utilized a generator to identify the most misleading transformations, aiming to increase the loss of the discriminator and optimize it in an adversarial manner. Dong et al. [40] employed minimax optimization to learn the distribution of potential adversarial examples, aiming to represent potential 'enemy' samples. Kim et al. [41] generated samples and fine-tuned the embedding function, focusing on densifying the embedding space around label samples, thereby enhancing the classifier's robustness during query handling.

In the context of rare skin disease classification tasks, where data availability is limited and significant intra-class variations exist, uncertainty during prototype construction poses a challenge. To address this, we propose employing hyperbolic space for precise estimation during training. By replacing prototypes with the highest uncertainty, we create adversarial prototypes. Integrating these prototypes into the training process enhances the model's discriminatory capabilities in exceptional cases and bolsters overall robustness.



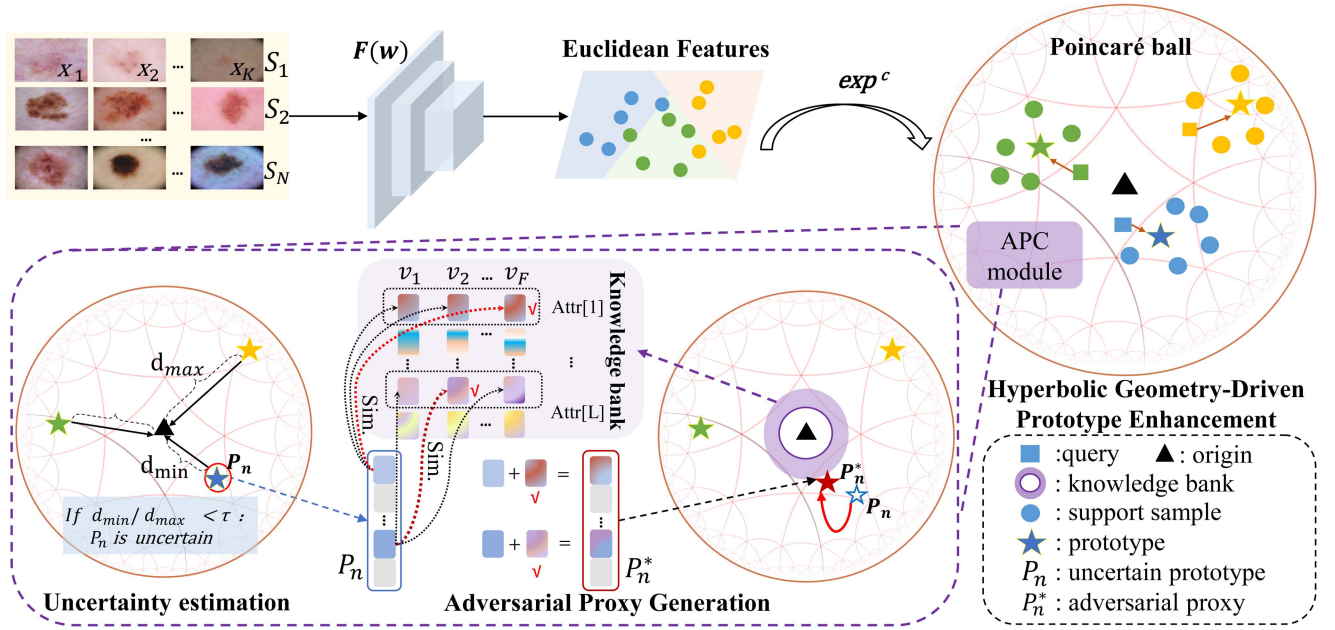


Fig. 2. Illustration of our HGRE framework. HGRE implements metric learning in the hyperbolic space, leveraging its unique structure and implicit hierarchical relationships to achieve superior metric learning outcomes. To further enhance the model's robustness against uncertain prototypes, an APC module is devised. This module first filters uncertain prototypes using hyperbolic geometry and then constructs adversarial proxies for uncertain prototypes during the training process.

### III. METHOD

Our study begins with a dataset  $\mathcal{D}$ , comprising pairs of images and their corresponding labels, denoted as  $\{\mathbf{x}_i, \mathbf{l}_i\}$ , where  $\mathbf{x}_i$  is an input image and  $\mathbf{l}_i$  is its one-hot encoded label. According to the customary protocols for FSL, we categorize the set of classes into base classes  $C_{base}$  and novel classes  $C_{novel}$ , ensuring there is no overlap. We partitioned  $C_{base}$  into a training set  $D_{train}$  and a validation set  $D_{val}$  to facilitate the model training and evaluation. Our goal is to train a model using  $D_{train}$  that can effectively generalize to  $C_{novel}$ , even with few-shot training examples. To achieve this, we simulate  $N$ -way  $K$ -shot learning scenarios, termed as episodes, from  $C_{base}$  during training. Each episode includes a support set  $S$ , with  $K$  samples per class for  $N$  classes, and a query set  $Q$  with  $T$  instances per class for the same  $N$  classes.

The proposed HGRE framework is designed to capitalize on the inherent hierarchical structures within classes, operating within the hyperbolic space to enhance metric learning. As depicted in Fig. 2, this framework is built around a feature extractor  $\mathcal{F}(w)$  responsible for converting input images into high-dimensional features. These features are then mapped from the Euclidean space  $\mathbb{R}^n$  onto the hyperbolic space  $\mathbb{D}_c^n$  using an exponential map  $\exp^c: \mathbb{R}^n \rightarrow \mathbb{D}_c^n$ . For each class, we calculate the class prototype by averaging the feature representation of the corresponding support samples. The APC module first filters uncertain prototype representations by evaluating their distance from the origin of the hyperbolic space. Then, for each uncertain prototype, the APC module constructs an adversarial proxy, which is used to strengthen the training process to increase the robustness of the model.

#### A. Hyperbolic Prototypical Network (H-ProtoNet)

Skin disease datasets often exhibit hierarchical structures in the label space, providing valuable prior knowledge about class relationships. Capturing this hierarchical structure is essential for distance metrics, particularly in tasks involving rare skin disease diagnosis where training data is scarce. However, most existing methods measure distance metrics based on Euclidean space, which inadequately represent these hierarchies [8], [9], [10]. To address this challenge, we propose conducting rare skin disease diagnosis in **hyperbolic space**, which outperforms Euclidean space in effectively representing hierarchical relationships [21], [22], [23]. By harnessing this representational capability, computing distance metrics in hyperbolic space enhances the representation of similarity between samples and classes. We utilize the  $n$ -dimensional Poincaré ball model  $\mathbb{D}_c^n = \{w \in \mathbb{R}^n : c\|w\|^2 < 1, c \geq 0\}$  to represent hyperbolic space, as stated in the literature [21], [23], [32]. In this model, the curvature parameter  $c$  moderates the compromise between Euclidean and hyperbolic properties. As  $c$  tends to 0, the space nature transitions to Euclidean.

Mapping features  $e_i$  from Euclidean space to the Poincaré ball  $\mathbb{D}_c^n$  requires bilateral conversion functions: the exponential map and its inverse, the logarithm map. The exponential mapping function  $\exp^c: \mathbb{R}^n \rightarrow \mathbb{D}_c^n$  is defined as follows:

$$\exp_w^c(e_i) = w \oplus_c (\tanh(\sqrt{c} \lambda_w^c \|e_i\| / 2) e_i / (\sqrt{c} \|e_i\|)). \quad (1)$$

Typically, the base point  $x$  is set to zero for mathematical simplicity, resulting in little negligible effect on results [46].

The hyperbolic space is not a vector space. To perform operations such as addition, we need introduce the so-called rotational

vector [47]. For  $x, y \in \mathbb{D}_c^n$  pairs, the Möbius addition operation  $\oplus_c$  is defined as [32]:

$$x \oplus_c y = \frac{(1 + 2c \langle x, y \rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c \langle x, y \rangle + c^2\|x\|^2\|y\|^2}. \quad (2)$$

And the hyperbolic distance between  $x, y$  is defined as:

$$d_c(x, y) = 2/\sqrt{c} \operatorname{arctanh}(\sqrt{c}\| -x \oplus_c y \|). \quad (3)$$

Our method mapped each Euclidean feature  $e_i$  into the hyperbolic space  $\mathbb{D}_c^n$  as  $h_i$  and classified query samples by assessing their hyperbolic distance to the hyperbolic prototype  $p_n$  of each class. Due to the hyperbolic space's adeptness at capturing hierarchical relationships, it is particularly suitable for metric learning. The subsequent APC module estimates prototype uncertainty by calculating the distance from this prototype to the hyperbolic space origin. The hyperbolic distance between  $h_i \in \mathbb{D}_c^n$  and  $p_n \in \mathbb{D}_c^n$  is defined as:

$$d_c(h_i, p_n) = 2/\sqrt{c} \operatorname{arctanh}(\sqrt{c}\| -h_i \oplus_c p_n \|). \quad (4)$$

Notably, under the limit of  $c$  approaching 0, it simplifies to the Euclidean distance  $d_c(h_i, p_n) = 2\|h_i - p_n\|$ .

Based on  $d_c$ , metric learning aims to assign a loss function that propels  $h_i$  toward its true class prototype  $p_n$  in hyperbolic space while simultaneously distancing it from other class prototypes  $p_{n'}$ . To this end, we define the following loss for the support set  $S_n = \{e_i | y_i = n\}$  of the  $n_{th}$  class

$$\mathcal{L}_{hyp} = - \sum_{e_i \in S_c} \log \frac{e^{-d(h_i, p_n)}}{\sum_{n' \in N} e^{-d(h_i, p_{n'})}}. \quad (5)$$

By minimizing this loss, we can directly update the feature extractor  $\mathcal{F}(w)$  in the hyperbolic space. In contrast to the feature extractor of ProtoNet [15] which is updated in the Euclidean space, our  $\mathcal{F}(w)$  can implicitly encode hierarchical relationships among different classes, thereby enhancing metric learning.

## B. APC Module

In the context of rare skin diseases, the limited availability of training data renders the model vulnerable to data uncertainty. The adversarial robust training (ART) theory [39], [40], [41], [45] enhances model robustness to uncertain data by first generating adversarial examples that elicit incorrect predictions from the model and then using these adversarial samples in training, enabling the model to accurately classify these challenging cases. By iterating this process, ART effectively fortifies the model against uncertain data. Building on this foundation, our APC module mitigates diagnostic errors caused by uncertain data through adversarial robust learning. It leverages the geometric properties of **hyperbolic space** to measure uncertainty and selects uncertain prototypes, which are then transformed into adversarial prototypes. These prototypes pose challenges during training, further bolstering the model's robustness against uncertain data. In constructing adversarial prototypes, it is crucial to not only confuse the model's prediction but also to ensure minimal alteration of the sample features. To achieve this, we implemented an adversarial prototype construction strategy

known as attribute obfuscation. This approach increases the complexity of the model's learning task while preserving its normal learning capacity.

**Uncertainty estimation:** Benefiting from the unique geometric properties of hyperbolic space, we utilize the geodesic distance to the hyperbolic origin as an elegant proxy to quantify feature uncertainty [21], [23]. Our strategy centers on identifying uncertain prototypes that require careful consideration during training. For each support set  $S_n = \{e_i | y_i = n\}$  in an  $N$ -way  $K$ -shot training episode, we calculate its hyperbolic prototype  $p_n$  by averaging the hyperbolic features of  $K$  support samples for each class. Across all classes in an episode, we calculate the geodesic distance from each prototype to the Poincaré ball origin and record the distance gap between the nearest and farthest prototypes. If this distance gap exceeds a predefined threshold  $\tau$ , the episode is deemed highly uncertain, and the nearest prototype to the Poincaré ball origin is labeled as uncertain.

**Adversarial Proxy Generation:** After filtering uncertain prototypes, the APC module strengthens the model's robustness to such prototypes by constructing an adversarial proxy for each one. These proxies are designed to complicate the training process, thus fortifying the model against uncertainty. The design of our APC module is inspired by the adversarial robust training (ART) theory [39], [40], [41], [45]. Many studies based on ART theory have shown that constructing adversarial samples and conducting adversarial training can improve the robustness of the model. The construction of adversarial samples should not only confuse the model's judgment but also ensure small changes in overall features.

Based on the ART theory, we construct adversarial prototypes by slightly altering specific attributes of the uncertain prototypes. These alterations serve two primary objectives.

First, we meticulously modulate the extent of the modifications to the original uncertain prototypes, to make model determination more difficult while preserving the integrity of their semantic content. Second, through this operation, the adversarial proxies is placed closer to the origin of the hyperbolic space, which increases the the uncertainty and intensify the training hurdles. This balance is crucial to maintain the prototype's relevancy while still contributing to the overall robustness of the model.

**Knowledge Bank Preparation:** The construction of adversarial proxies requires the preparation of a knowledge bank that accommodates multiple shared representations for each attribute across classes. Firstly, in order to ensure the diversity of representations in the knowledge bank, we randomly divide the features of all training samples into  $F$  subsets in each training epoch. We average the features within each subset, resulting in  $F$  general embeddings that form the initial knowledge bank  $B^i = \{v_1^i, \dots, v_F^i\}$  of the  $i_{th}$  epoch. Next, in order to further ensure that each general embedding encapsulates the common knowledge from multiple samples, we dynamically updated the knowledge bank  $B^i$  by calculating the weighted average of the embeddings from five consecutive epochs, i.e.,  $v_F^i = w_{i-4} \times v_{i-4}^{i-4} + w_{i-3} \times$

$v_F^{i-3} + w_{i-2} \times v_F^{i-2} + w_{i-1} \times v_F^{i-1} + w_i \times v_F^i$ . The weight values  $\{w_{i-4}, \dots, w_i\}$  for five epochs are empirically set to  $\{0.1, 0.1, 0.1, 0.2, 0.5\}$ , and considering that the features learned in the latest epoch are the most accurate, we assign the highest weight to the embeddings from the last epoch.

To further process the general embedding within the updated knowledge bank, each one is partitioned into  $L$  segments along the channel dimension, with each segment delineating a semantic attribute. The resultant average general embedding not only shows the general information of the same position attributes, but also significantly elevates the level of uncertainty. Affected by the hyperbolic geometrical properties, the general embeddings with high uncertainty will be closer to the Poincaré ball origin. Similarly, the integration of general embeddings and uncertain prototypes can not only increase the uncertainty of the prototypes, but also introduce the general information of the corresponding attributes to achieve a mild and effective adversarial proxy generation. By utilizing this multi-class average KB, we can leverage knowledgeable attributes for subsequent attribute obfuscation operations.

**Attribute Obfuscating:** Following the establishment of a knowledge bank, which contains the latest class-general knowledge, we utilize it to modify each uncertain prototype. For a given uncertain prototype  $p_n$ , we split it into  $L$  segments along the channel dimension and obtain  $L$  attributes.

For the  $F$  general embeddings, we also divide each of them  $L$  segments, so that each position has  $F$  attributes to choose from for obfuscation. A subset of  $M$  attributes in prototype  $p_n$  is then selected at random for deliberate obfuscation. For each chosen attribute  $p_n[j] \{p_n[1], \dots, p_n[M]\}$ , we identify the most similar attribute  $attr[j]$  among the  $F$  attributes in the corresponding position in the  $F$  general embeddings which stored in the knowledge bank. The obfuscation process is mathematically expressed as:

$$p_n^*[j] = \lambda_1 p_n[j] + \lambda_2 attr[j]. \quad (6)$$

By applying this process to all selected attributes, we generate an adversarial proxy  $p_n^*$  for the uncertain prototype  $p_n$ . Given that attributes from the knowledge bank inherently lie nearer to the Poincaré ball origin than those of  $p_n$ , the adversarial proxy  $p_n^*$  is consequently positioned closer to the origin, inherently embedding a greater degree of uncertainty. This characteristic intensifies the training difficulty, hence boosting the model's robustness to novel and uncertain prototypes.

Empirical studies highlight the pivotal role of the knowledge bank in this context of adversarial proxy creation. Substituting the knowledge bank's attributes with randomly generated equivalents does not translate to substantial enhancements in performance. This is attributed to the random attributes' inability to encapsulate relevant general information and to the excess perturbation they introduce, which potentially compromises the prototype's significant dynamics. Therefore, the employment of a knowledge bank replete with diverse class insights is essential to consistently elevate model proficiency via strategic attribute obfuscation.

## IV. EXPERIMENTS

### A. Datasets:

**The SD-198 dataset** [53] comprises 6,584 images from 198 fine-grained classes of skin diseases. These images were captured using various digital cameras and mobile devices. They have an original size of  $1640 \times 1130$  and were resized to  $224 \times 224$  for efficient training. We divided the dataset into training, validation, and test sets based on the sample number of each class. Specifically, we allocated the 98 classes with the largest number of samples to the training set, reserved 30 classes for validation, and designated the remaining 70 classes for testing. It should be noted that each class within the test database, representing rarer disease types, includes fewer than 20 images. Each base class, including both training and validation groups, possesses a maximum of 60 images only.

**The ISIC 2019 Skin Lesion dataset** [54], [55], [56] consists of 25,331 images of skin lesions associated with eight categories of skin conditions: melanocytic nevus (12,875), benign keratosis (2,624), melanoma (4,522), basal cell carcinoma (3,323), actinic keratosis (867), dermatofibroma (239), squamous cell carcinoma (628), and vascular lesion (253). Adhering to the few-shot classification protocol in [8], we utilized the four most populous classes for training purposes, while the remaining four classes were allotted for testing.

### B. Implementation Details

The proposed HGRE framework was implemented using the Pytorch platform, leveraging an NVIDIA GeForce RTX 2080Ti GPU to accelerate computational processes. In the context of FSL, it is imperative to effectively transfer knowledge from the base classes to novel classes that are represented by limited training samples. Overly deep networks have a propensity to overfit the base classes, which undermines their effectiveness in adapting to novel classes. In most FSL applications, architectures such as ResNet12 [57] and ConvNet are used as model backbone. However, our subsequent experiments show that for the skin Disease dataset, the feature extraction ability of ConvNet is insufficient, which limits the performance of the final model. Therefore, we have adopted ResNet12 as the backbone for our proposed method and other comparison methods. The optimization was conducted via the Adam algorithm with a learning rate at 0.005. The variability of training data was enriched through the application of online data augmentation strategies, such as random cropping, random affine transformations, and both vertical and horizontal flips. The training spanned 200 epochs, with a batch capacity of 200. In the process, each feature vector  $e_i$  was mapped onto a 512-dimensional Poincaré ball model. The hyperbolic space curvature  $c$  was set to 0.001 for 1-shot settings and 0.005 for 5-shot settings. To determine uncertain prototypes during training, we calculated the distance from each prototype to the origin within each batch. If the ratio of the minimum distance to the maximum distance fell below the pre-determined threshold  $\tau$  (set to 0.95 in our experiments), we considered the prototype closest to the origin as uncertain



**TABLE I**  
PERFORMANCE COMPARISON OF OUR HGRE FRAMEWORK WITH NINE COMPETING METHODS ON SD-198 AND ISIC 2019 DATASETS

Methods	SD-198 (4-way-1-shot)		SD-198 (4-way-5-shot)		ISIC 2019 (2-way-1-shot)		ISIC 2019 (2-way-5-shot)	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
ProtoNet (NIPS2017) [40]	69.62±0.34	89.65±0.14	85.56±0.09	96.93±0.13	53.62±0.61	55.67±0.73	60.10±0.53	64.37±0.73
LRDC (ICLR2021) [52]	69.84±1.13	88.69±0.58	86.17±0.74	97.02±0.20	53.77±0.38	55.69±0.63	57.93±0.31	61.22±0.58
DeepBDC (CVPR2022) [49]	60.50±0.66	83.23±1.63	82.20±0.58	95.58±0.07	55.69±0.58	59.15±0.77	63.98±1.05	69.79±1.21
RENet (ICCV2021) [17]	66.16±0.99	87.41±0.56	70.76±0.78	89.78±0.30	55.08±0.75	58.11±1.23	63.76±0.74	69.85±0.58
FRN (CVPR2021) [47]	68.55±0.36	88.76±0.15	85.45±0.15	95.59±0.05	56.64±0.33	59.53±0.66	63.44±0.30	69.73±0.47
CPN (AAAI2023) [27]	58.51±0.94	82.31±0.56	76.53±0.58	93.41±0.43	58.19±0.21	61.92±0.36	65.10±0.59	71.16±0.80
Bi-FRN (AAAI2023) [48]	70.52±0.24	80.49±0.14	86.22±0.14	90.88±0.09	56.93±0.35	56.97±0.36	64.81±0.31	64.84±0.30
AFRN (PR2024) [24]	70.86±0.31	89.22±0.10	83.19±0.19	94.99±0.07	<b>59.24±0.30</b>	<b>63.55±0.58</b>	64.96±0.45	72.09±0.37
HFA (NIPS2022) [13]	67.89±0.24	89.48±0.16	86.58±0.20	97.23±0.07	57.62±0.24	62.12±0.46	64.87±0.38	70.76±0.52
<b>HGRE(ours)</b>	<b>71.37±0.13</b>	<b>90.96±0.04</b>	<b>86.69±0.24</b>	<b>97.41±0.08</b>	58.06±0.30	63.18±0.26	<b>67.11±0.67</b>	<b>74.58±0.77</b>

Performance is assessed by ACC (%) and AUC (%). All results represent the average of five independent runs, with the best and second-best results in each column highlighted in bold and underlined.

and obfuscated it using the knowledge bank. For the remaining hyper-parameters, ( $M = 3$ ) out of  $L = 10$  attributes of each prototype were selected for obfuscation, and the selected attributes and the attributes from the knowledge bank were combined with the weights of  $\lambda_1 = 0.9$  and  $\lambda_2 = 0.1$  to generate an adversarial proxy.

To ensure reliability, we performed quintuplicate experiments for each method and averaged the results to provide a robust assessment. The evaluation metrics used for this study include accuracy (ACC) and the area under the receiver operating characteristic curve (AUC). The reported results reflect the average performance across over 2000 batches in the inference stage.

### C. Comparing to Existing Methods

We conducted a comparative evaluation of the proposed HGRE framework against nine state-of-the-art FSL methods. Since HGRE is based on metric learning, competing methods include seven metric learning models. *ProtoNet* [15] focuses on learning a metric space that enables classification based on the distance to the prototype representation of each class. *DeepBDC* [17] concurrently deliberates both marginal and joint distributions to learn image representation via assessing the discrepancy between the joint characteristic functions of embedded features and the product of their marginals. *FRN* [48] reconceptualizes few-shot classification as a reconstruction task in a latent space, whereby the query feature map is reconstructed by the network using the supporting features of a given class, allowing prediction of the query's class affiliation. *RENet* [18] exploits relation-based patterns within and across images through a self-correlational representation framework and a cross-correlational attention mechanism. *CPN* [49] integrates metric learning with semantic information by constructing compositional prototypes based on human-annotated attributes. *Bi-FRN* [50] employs bi-reconstruction mechanisms to explore subtle intra-class and inter-class variations, enhancing feature discriminability. *AFRN* [51] modifies the inter-class score in TDM and incorporates a center loss to FRN to amplify inter-class discrepancies. Meanwhile, competing methods also include two solutions that utilize data augmentation. *LRDC* [16] aims to calibrate the distribution of sample classes by utilizing class transfer statistics. It subsequently generates additional samples from the

calibrated distribution to augment the input for the classifier. *HFA* [52] employs hyperbolic normal distributions and neural ordinary differential equations to model augmented features, leveraging geometric properties for enhanced performance.

We re-implemented these methods using the hyper-parameter settings reported in their original papers. This ensured a fair and standardized evaluation on the same data split as utilized in our HGRE framework.

1) *Results on SD-198 Dataset*: To assess the efficacy of our proposed HGRE framework with varying numbers of support images, we conducted experiments on the SD-198 dataset using the 4-way-1-shot and 4-way-5-shot settings. The results, as evidenced in Table I, demonstrate that our HGRE consistently exhibits superior performance under both settings. The experimental results on the SD-198 dataset highlight the challenges encountered in real-world clinical scenarios, primarily due to its limited sample number (e.g., the largest base class contains only 60 images) and significant intra-class variation. This scarcity adversely affects the performance of many competing methods. For instance, CPN [49] demonstrates notably poorer results on the SD-198 dataset than on the ISIC 2019 dataset. Additionally, the considerable variation among images within the same disease category exacerbates the difficulties presented by this dataset. In such circumstances, the straightforward averaging of support images to create prototypes can result in ambiguous representations, leading to classification errors and hindering performance improvements. For example, RENet [18] exhibits a notable decline in performance in the 5-shot setting compared to the 1-shot setting, underscoring the detrimental impact of intra-class diversity on metric-based methods. In contrast, our HGRE framework leverages implicit hierarchical class relationships and employs uncertainty-based adversarial training, significantly enhancing the classification performance and model robustness. Even in scenarios with limited sample sizes and high intra-class variability, our method consistently achieves superior performance across all metrics and settings.

2) *Results on ISIC 2019 Dataset*: To assess the generalization capabilities of our HGRE framework across different datasets, we conducted a parallel set of experiments on the ISIC 2019 dataset using the 2-way-1-shot and 2-way-5-shot settings. The results of each method are presented in Table I. It shows that under these two settings, our HGRE has achieved excellent

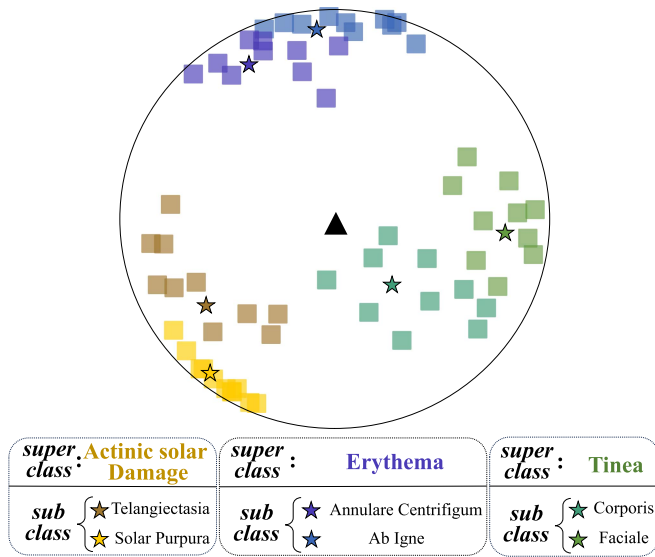


Fig. 3. HGRE embeddings of classes in the SD-198 dataset (train set) on the Poincaré disc. Each square within the disc represents a query sample, with distinct colors indicating different classes. The pentagram within the same color signifies the feature prototype for that class, which is computed as the average of the 5 support samples (6-way-5-shot).

performance. Unlike the SD-198 dataset, the ISIC 2019 dataset features fewer classes but a larger number of samples per class, particularly in the base classes. The main challenge of the ISIC 2019 dataset lies in its significant intra-class variance; however, the increased sample size diverges significantly from the clinical scenarios typically encountered in rare disease diagnoses. Although AFRN AFRN [51] achieves a 1.18% improvement in ACC over our HGRE framework in the 1-shot setting, it is important to note that in the 5-shot setting, which better reflects real-world scenarios, our method surpassed AFRN by 2.15% in ACC and 2.49% in AUC. Additionally, our HGRE framework exhibits significant performance improvements over other competing methods. These results highlight the superior robustness of our HGRE framework in managing intra-class variability and data uncertainty.

#### D. Visualization Analysis

We employed t-SNE visual mappings to gain insights into the data distribution of the SD-198 dataset in the hyperbolic space (the Poincaré disc). Fig. 3 presents a snapshot of the data projection. The graphic illustrates superclass directionality divergences and subclass aggregation within these groupings, consistent with the hyperbolic hierarchy. At the same time, the geodesic distance metric, inherent to hyperbolic space, ensures that subclasses belonging to the same superclass maintain a degree of separability in the feature space. Furthermore, the feature embedding distribution exhibits a characteristic behavior: prototypes of classes with tightly clustered features (low uncertainty) are located near the periphery, whereas prototypes of classes with more scattered features (high uncertainty) converge towards the origin. This observation aligns with our emphasis

on accurate prototype estimation and uncertainty estimation in the hyperbolic space.

#### E. Ablation Study

To further validate the efficacy of integrating hyperbolic geometry and adversarial robust training into FSL, we conducted ablation studies on two datasets under the 1-shot and 5-shot settings. The results of each variant are summarized in Table II. ProtoNet [15] utilizes neither hyperbolic metric learning nor the APC module, serving as the baseline model. The H-ProtoNet variant, integrating hyperbolic metric learning within ProtoNet, consistently outperforms the baseline, particularly in terms of AUC, suggesting that the metric learning in the hyperbolic space exhibits superior performance compared to that in the Euclidean space. Further deploying the HGRE framework, with the APC module to fortify against uncertain prototypes, we noted its superior performance over H-ProtoNet. It illustrates the APC module's proficiency in incorporating uncertainty into adversarial training, enhancing the network's robustness to uncertain samples and ultimately leading to improved classification performance. These findings underscore the significance of exploring uncertainty robustness in FSL for enhancing rare skin disease diagnosis accuracy.

### V. DISCUSSION

#### A. Hyper-Parameter Settings

Given the findings from our previous experiments, we narrowed our focus to the SD-198 dataset with the 1-shot setting to assess the impact of hyperparameters on the performance of our HGRE framework.

**Obfuscation Intensity:** To determine the optimal obfuscation intensity, we evaluated varying proportions and frequencies of obfuscation applications and their impact on overall performance. Specifically, we calculated the distance from each prototype to the Poincaré ball origin and analyzed the ratio between the nearest and farthest prototypes. A smaller ratio, falling beneath a preset threshold  $\tau$ , triggered obfuscation of the nearest prototype via the knowledge bank. A larger  $\tau$  leads to increased obfuscation frequency. For obfuscation,  $M$  out of  $L$  attributes of each prototype were randomly selected for obfuscation. These selected attributes were then weighted alongside the attributes from the knowledge bank at predefined proportions ( $\lambda_1$  and  $\lambda_2$ ). Table III presents the experimental results, indicating that neither the threshold nor the obfuscation proportion significantly impacted the results. Across various intensity settings, our method consistently outperforms comparative methods, leading us to adopt an optimal set of parameters ( $\tau=0.95$ ,  $L=10$ ,  $M=3$ ,  $\lambda_1=0.9$ , and  $\lambda_2=0.1$ ) as the default configuration for our experiments.

**Curvature  $c$ :** We evaluated the curvature  $c$  within the range of 0.1 to  $1e-6$ , with results summarized in Table IV. Our method exhibited a specific requirement for  $c$ , though it was not overly stringent. Notably, when  $c$  fell within the range of  $1e-2$  to  $1e-5$ , the model's performance was deemed satisfactory. As  $c$  approaching zero, indicating an approximation to Euclidean



**TABLE II**  
ABLATION STUDY ON THE SD-198 AND THE ISIC 2019 DATASETS

Hyperbolic	APC	SD-198 (4-way-1-shot)		SD-198 (4-way-5-shot)		ISIC 2019 (2-way-1-shot)		ISIC 2019 (2-way-5-shot)	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
		69.62±0.34	89.65±0.14	85.56±0.09	96.93±0.13	53.62±0.61	55.67±0.73	60.10±0.53	64.37±0.73
✓		69.70±0.07	90.50±0.08	85.97±0.18	97.22±0.10	55.54±0.50	59.18±0.79	64.60±0.52	71.34±1.11
✓	✓	<b>71.37±0.13</b>	<b>90.96±0.04</b>	<b>86.69±0.24</b>	<b>97.41±0.08</b>	<b>58.06±0.30</b>	<b>63.18±0.26</b>	<b>67.11±0.67</b>	<b>74.58±0.77</b>

It presents the mean±standard deviation of performance metrics (%), evaluated on the SD-198 dataset under the 4-way-1-shot and the 4-way-5-shot settings, as well as the ISIC 2019 dataset under the 2-way-1-shot and the 2-way-5-shot settings. The best and second-best results were highlighted in bold and underlined.

**TABLE III**  
ABLATION STUDY CONDUCTED ON THE SD-198 DATASET USING THE 4WAY-1-SHOT SETTING WITH DIFFERENT THRESHOLD  $\tau$  AND PARAMETERS  $L$ ,  $M$ ,  $\lambda_1$ ,  $\lambda_2$  FOR OBFUSCATION

$L$ and $M$	$\lambda_1$ and $\lambda_2$	$\tau=0.9$		$\tau=0.925$		$\tau=0.95$		$\tau=0.975$	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
10:1	0.9:0.1	70.86±0.11	90.78±0.09	71.21±0.20	91.00±0.12	70.76±0.33	90.67±0.10	71.04±0.25	90.78±0.09
10:5	0.9:0.1	70.76±0.27	90.62±0.15	70.35±0.15	90.51±0.08	70.53±0.32	90.43±0.07	70.67±0.38	90.54±0.12
10:3	0.9:0.1	70.78±0.22	90.61±0.09	70.84±0.21	90.55±0.11	<b>71.37±0.13</b>	<b>90.96±0.04</b>	70.82±0.18	90.55±0.09
10:3	0.95:0.05	70.44±0.26	90.46±0.10	70.47±0.31	90.67±0.06	70.79±0.34	90.78±0.14	70.62±0.44	90.61±0.08
10:3	0.8:0.2	70.64±0.44	90.61±0.05	70.40±0.25	90.46±0.13	70.54±0.29	90.56±0.10	70.60±0.07	90.64±0.12

It presents the mean±standard deviation of performance metrics (%). The best results were highlighted in bold.

**TABLE IV**  
ABLATION STUDY CONDUCTED ON THE SD-198 DATASET USING THE 4WAY-1-SHOT SETTING WITH DIFFERENT CURVATURE  $c$

Dataset	Curvature $c$	SD-198 (4-way-1-shot)	
		ACC	AUC
SD-198	1E-1	63.37±0.37	85.85±0.27
	1E-2	70.31±0.23	90.67±0.07
	1E-3	<b>71.37±0.13</b>	<b>90.96±0.04</b>
	1E-4	<u>71.28±0.45</u>	<u>90.78±0.09</u>
	1E-5	70.84±0.20	90.51±0.10
	1E-6	69.89±0.20	89.91±0.06

The best and second-best results were highlighted in bold and underlined.

space, and intriguingly, at  $c = 1e - 6$ , our model's performance was comparable to that of ProtoNet in the Euclidean space [15]. Excessive curvature ( $c = 1e - 1$ ), in contrast, led to a significant decline in performance. Therefore, we opted for  $c = 0.001$  for the 1-shot setting and  $c = 0.005$  for the 5-shot setting.

## B. Origin Random Sampling Vs. Knowledge Bank

To demonstrate the essential role of the knowledge bank, we conducted a comparative evaluation with and without the knowledge bank in the APC module. Table V details the results of substituting the knowledge bank with randomly sampled pseudo information from the hyperbolic space's origin. The experimental outcomes indicate that this altered HGRE iteration (*i.e.*, HGRE with Pseudo KB) cannot achieve notable performance improvements over H-ProtoNet, affirming the indispensable role of an authentic knowledge bank – stocked with generalized information – in adversarial robust learning.

## C. Different Backbones

We extended our analysis to diverse backbones under the 5-shot setting, across two datasets, to probe the robustness of integrating hyperbolic geometry and APC module. The results of each variant are summarized in Fig 4. It shows that, due to the

less potent feature extraction capabilities of ConvNet, the performance improvement of our method using ConvNet backbone is relatively less than that using ResNet12 backbone on SD-198 dataset. However, on the more challenging ISIC 2019 dataset, our method demonstrates marked improvements regardless of the backbone. Consequently, irrespective of dataset or backbone variations, the harmonious integration of hyperbolic space and the APC module consistently yields stable performance gains.

## D. Limitations and Future Work

First, due to the challenges of obtaining genuine samples of rare diseases, we simulate diagnostic scenarios by sampling from public datasets. However, newly collected images of rare diseases often display significant quality differences from existing images, influenced by variations in hospitals, skin conditions, and skin tones. While our HGRE method enhances model robustness through adversarial training, it may still experience performance degradation when confronted with these domain discrepancies. In our future work, we plan to incorporate cross-domain adaptation into HGRE to improve its generalization and transferability, allowing it to adapt more effectively to different data distributions in novel classes.

Second, this study's hierarchical knowledge acquisition regarding rare diseases is primarily based on image data. However, textual data related to rare diseases, such as disease names and corresponding descriptions, can also be utilized. An intriguing direction for future research is to explore language-enhanced hierarchy analysis for rare disease diagnosis, which aims to enhance the model's knowledge acquisition by leveraging the text embedding capabilities of large language models.

Third, in clinical practice, some subjects may present with unseen rare diseases that were never included in the training dataset. Accurately classifying these unseen cases poses a significant challenge. Thus, another promising area for future research is to investigate rare disease diagnosis within the framework of zero-shot learning.

TABLE V

MEAN  $\pm$  STANDARD DEVIATION OF PERFORMANCE METRICS (%), OBTAINED FROM AN ABLATION STUDY CONDUCTED ON THE SD-198 DATASET WITH THE ORIGIN RANDOM KNOWLEDGE BANK SETTING

Dataset	Methods	4-way-1-shot		4-way-5-shot	
		ACC	AUC	ACC	AUC
SD-198	ProtoNet	69.62 $\pm$ 0.34	89.65 $\pm$ 0.14	85.56 $\pm$ 0.09	96.93 $\pm$ 0.13
	H-ProtoNet	69.70 $\pm$ 0.07	90.50 $\pm$ 0.08	85.97 $\pm$ 0.18	97.22 $\pm$ 0.10
	HGRE(Pseudo KB)	70.17 $\pm$ 0.23	90.33 $\pm$ 0.09	86.27 $\pm$ 0.25	97.34 $\pm$ 0.12
	HGRE	<b>71.37<math>\pm</math>0.13</b>	<b>90.96<math>\pm</math>0.04</b>	<b>86.69<math>\pm</math>0.24</b>	<b>97.41<math>\pm</math>0.08</b>

The best and second-best results were highlighted in bold and underlined.

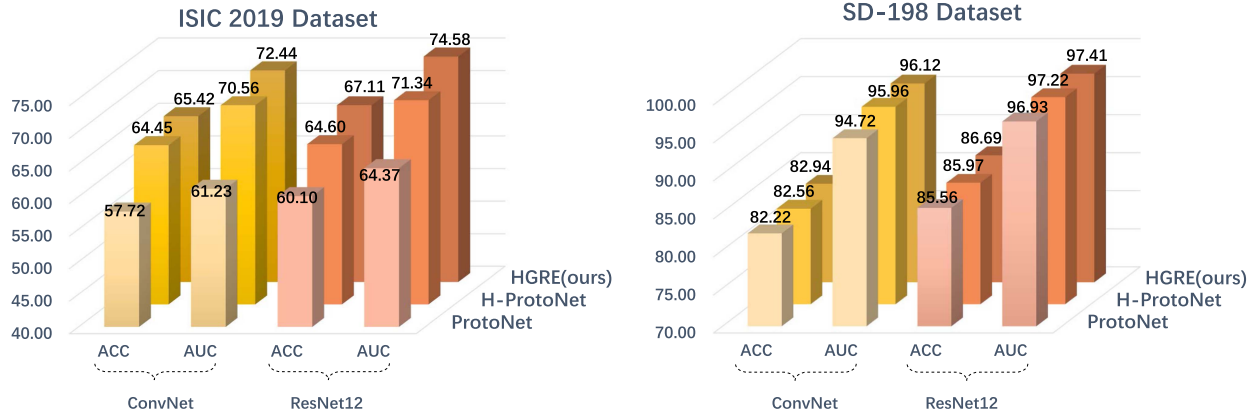


Fig. 4. Results of ablation study on the ISIC 2019 and the SD-198 datasets with different backbones under the 2-way-5-shot and the 4-way-5-shot setting.

## VI. CONCLUSION

In this study, we introduce the HGRE framework for rare skin disease diagnosis, leveraging two hyperbolic-geometry properties to create an enhanced FSL-based diagnosis model. Our HGRE initiates metric learning in the hyperbolic space, capitalizing on its inherent implicit class hierarchical structures. Then it couples this with the APC module that sieves uncertain prototypes according to their distance to the hyperbolic origin, thereby enhancing the model's robustness to uncertain prototypes by constructing adversarial proxies. Leveraging the two unique geometrical properties of hyperbolic space, our unified and coherent framework effectively addresses the limitations of insufficient hierarchical relation utilization and prototype uncertainty in FSL-based rare disease diagnosis. Extensive experiments on two skin lesion classification datasets demonstrate the superior performance of the proposed HGRE framework, highlighting its potential for accurate and robust rare skin disease diagnosis.

## REFERENCES

- [1] Y. Liu et al., "A deep learning system for differential diagnosis of skin diseases," *Nat. Med.*, vol. 26, no. 6, pp. 900–908, 2020.
- [2] Z. Liu, R. Xiong, and T. Jiang, "CI-Net: Clinical-inspired network for automated skin lesion recognition," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 619–632, Mar. 2023.
- [3] G. Yue, P. Wei, T. Zhou, Q. Jiang, W. Yan, and T. Wang, "Toward multicenter skin lesion classification using deep neural network with adaptively weighted balance loss," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 119–131, Jan. 2023.
- [4] X. Xing et al., "Categorical relation-preserving contrastive knowledge distillation for medical image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2021, pp. 163–173.
- [5] A. Paul, Y. Tang, T. C. Shen, and R. M. Summers, "Discriminative ensemble learning for few-shot chest X-ray diagnosis," *Med. Image Anal.*, vol. 68, 2021, Art. no. 101911.
- [6] S. Chao and D. Belanger, "Generalizing few-shot classification of whole-genome doubling across cancer types," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 3382–3392.
- [7] S. Mai, Q. Li, Q. Zhao, and M. Gao, "Few-shot transfer learning for hereditary retinal diseases recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2021, pp. 97–107.
- [8] 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. Med. Image Comput. Computer Assist. Intervention*, 2020, pp. 357–366.
- [9] 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, 2021, Art. no. 107797.
- [10] J. Sun, D. Wei, K. Ma, L. Wang, and Y. Zheng, "Unsupervised representation learning meets pseudo-label supervised self-distillation: A new approach to rare disease classification," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2021, pp. 519–529.
- [11] Z. Zhang, C. Lan, W. Zeng, Z. Chen, and S.-F. Chang, "Uncertainty-aware few-shot image classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021.
- [12] J. Xu and H. Le, "Generating representative samples for few-shot classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9003–9013.
- [13] Y. Cao et al., "Few-shot object detection via association and discrimination," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 16570–16581.
- [14] B. Li, B. Yang, C. Liu, F. Liu, R. Ji, and Q. Ye, "Beyond max-margin: Class margin equilibrium for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7363–7372.
- [15] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4080–4090.
- [16] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *Proc. Int. Conf. Learn. Representations*, 2021.
- [17] J. Xie, F. Long, J. Lv, Q. Wang, and P. Li, "Joint distribution matters: Deep Brownian distance covariance for few-shot classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7972–7981.
- [18] D. Kang, H. Kwon, J. Min, and M. Cho, "Relational embedding for few-shot classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8822–8833.

- [19] H. Cheng, S. Yang, J. T. Zhou, L. Guo, and B. Wen, "Frequency guidance matters in few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11814–11824.
- [20] L. Sun, M. Zhang, B. Wang, and P. Tiwari, "Few-shot class-incremental learning for medical time series classification," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 4, pp. 1872–1882, Apr. 2024.
- [21] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky, "Hyperbolic image embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6418–6428.
- [22] Z. Yu et al., "Skin lesion recognition with class-hierarchy regularized hyperbolic embeddings," in *Proc. Med. Image Comput. Comput. Assist. Intervention*, 2022, pp. 594–603.
- [23] M. G. Atigh, J. Schoep, E. Acar, N. Van Noord, and P. Mettes, "Hyperbolic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4453–4462.
- [24] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.
- [25] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv:1803.02999*.
- [26] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Representations*, 2016.
- [27] A. A. Rusu et al., "Meta-learning with latent embedding optimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [28] O. Vinyals et al., "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29.
- [29] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31.
- [30] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7260–7268.
- [31] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [32] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.
- [33] M. Nickel and D. Kiela, "Learning continuous hierarchies in the lorentz model of hyperbolic geometry," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3779–3788.
- [34] F. Sala, C. De Sa, A. Gu, and C. Ré, "Representation tradeoffs for hyperbolic embeddings," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4460–4469.
- [35] R. Sarkar, "Low distortion delaunay embedding of trees in hyperbolic plane," in *Proc. Int. Symp. Graph Drawing*, 2011, pp. 355–366.
- [36] B. Chen, W. Peng, X. Cao, and J. Röning, "Hyperbolic uncertainty aware semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1275–1290, Feb. 2024.
- [37] A. Montanaro, D. Valsesia, and E. Magli, "Rethinking the compositionality of point clouds through regularization in the hyperbolic space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 33741–33753.
- [38] R. Ma, P. Fang, T. Drummond, and M. Harandi, "Adaptive poincaré point to set distance for few-shot classification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1926–1934.
- [39] Y. Bin et al., "Adversarial semantic data augmentation for human pose estimation," in *Proc. Computer Vis.–ECCV 2020, 16th Eur. Conf.*, 2020, pp. 606–622.
- [40] Y. Dong, Z. Deng, T. Pang, J. Zhu, and H. Su, "Adversarial distributional training for robust deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 8270–8283.
- [41] J. Kim, H. Kim, and G. Kim, "Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning," in *Proc. Computer Vis.–ECCV 2020, 16th Eur. Conf.*, 2020, pp. 599–617.
- [42] S. Huang, Z. Lu, K. Deb, and V. N. Boddeti, "Revisiting residual networks for adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8202–8211.
- [43] L. Fan, S. Liu, P.-Y. Chen, G. Zhang, and C. Gan, "When does contrastive learning preserve adversarial robustness from pretraining to finetuning?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 21480–21492.
- [44] Q. Zeng, Y. Xie, Z. Lu, and Y. Xia, "PEFAT: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15671–15680.
- [45] K. Li, Y. Zhang, K. Li, and Y. Fu, "Adversarial feature hallucination networks for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13470–13479.
- [46] A. Ermolov, L. Mirvakhabova, V. Khrulkov, N. Sebe, and I. Oseledets, "Hyperbolic vision transformers: Combining improvements in metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7409–7419.
- [47] A. Ungar, "A gyrovector space approach to hyperbolic geometry," 2022.
- [48] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8012–8021.
- [49] Q. Lyu and W. Wang, "Compositional prototypical networks for few-shot classification," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 7, pp. 9011–9019.
- [50] J. Wu et al., "Bi-directional feature reconstruction network for fine-grained few-shot image classification," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 3, pp. 2821–2829.
- [51] X. Li, Z. Guo, R. Zhu, Z. Ma, J. Guo, and J.-H. Xue, "A simple scheme to amplify inter-class discrepancy for improving few-shot fine-grained image classification," *Pattern Recognit.*, vol. 156, 2024, Art. no. 110736.
- [52] Z. Gao, Y. Wu, Y. Jia, and M. Harandi, "Hyperbolic feature augmentation via distribution estimation and infinite sampling on manifolds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 34421–34435.
- [53] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *Proc. Computer Vis.–ECCV 2016, 14th Eur. Conf.*, 2016, pp. 206–222.
- [54] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.
- [55] N. C. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, 2018, pp. 168–172.
- [56] H. P. Carlos et al., "BCN20000: Dermoscopic lesions in the wild," *Sci. Data*, vol. 11, no. 1, p. 641, 2024.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.