Hyperbolic Self-Paced Multi-Expert Network for Cross-Domain Few-Shot Facial Expression Recognition

Xueting Chen, Yan Yan, *Senior Member, IEEE*, Jing-Hao Xue, *Senior Member, IEEE,* Chang Shu, and Hanzi Wang, *Senior Member, IEEE*

***Abstract*—Recently, cross-domain few-shot facial expression recognition (CF-FER), which identifies novel compound expressions with a few images in the target domain by using the model trained only on basic expressions in the source domain, has attracted increasing attention. Generally, existing CF-FER methods leverage the multi-dataset to increase the diversity of the source domain and alleviate the discrepancy between the source and target domains. However, these methods learn feature embeddings in the Euclidean space without considering imbalanced expression categories and imbalanced sample difficulty in the multi-dataset. Such a way makes the model difficult to capture hierarchical relationships of facial expressions, resulting in inferior transferable representations. To address these issues, we propose a hyperbolic self-paced multi-expert network (HSM-Net), which contains multiple mixture-of-experts (MoE) layers located in the hyperbolic space, for CF-FER. Specifically, HSM-Net collaboratively trains multiple experts in a self-distillation manner, where each expert focuses on learning a subset of expression categories from the multi-dataset. Based on this, we introduce a hyperbolic self-paced learning (HSL) strategy that exploits sample difficulty to adaptively train the model from easy-to-hard samples, greatly reducing the influence of imbalanced expression categories and imbalanced sample difficulty. Our HSM-Net can effectively model rich hierarchical relationships of facial expressions and obtain a highly transferable feature space. Extensive experiments on both in-the-lab and in-the-wild compound expression datasets demonstrate the superiority of our proposed method over several state-of-the-art methods. Code will be released at https://github.com/cxtjl/HSM-Net.**

***Index Terms*—Compound facial expression recognition, Cross-domain few-shot learning, Self-paced learning, Mixture-of-experts, Hierarchical representation learning.**

This work was supported by the National Key Research and Development Program of China under Grant 2022ZD0160402, the National Natural Science Foundation of China under Grant 62372388 and Grant U21A20514, the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City under Grant 3502Z20241029 and Grant 3502Z20241027, and the Fundamental Research Funds for the Central Universities under Grant 20720240076 and Grant ZYGX2021J004. *(Corresponding author: Yan Yan.)*

X. Chen and Y. Yan are with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China and the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: xtchen@stu.xmu.edu.cn; yanyan@xmu.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (e-mail: jinghao.xue@ucl.ac.uk).

C. Shu is with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: changshu@uestc.edu.cn).

Hanzi Wang is with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China and the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 201210, China (e-mail: hanzi.wang@xmu.edu.cn).

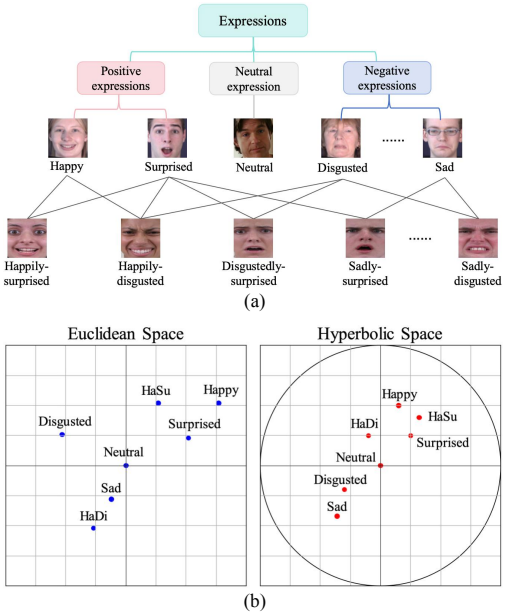


Fig. 1: Illustration of (a) hierarchical relationships of facial expressions and (b) feature distributions in the Euclidean and hyperbolic spaces, where 'HaSu' and 'HaDi' represent the happily-surprised and happily-disgusted expressions, respectively. The hierarchy arises from compound expressions (combinations of two basic expressions), basic expressions, and emotional valence (positive, neutral, and negative expressions).

I. Introduction

FACIAL expressions, as one of the most natural and universal ways for humans to express their emotions, play an important role in interpersonal communication [1], [2]. Over the past few decades, facial expression recognition (FER) has attracted considerable attention in computer vision due to its widespread applications, such as human-computer interaction, psychological assessment, and interactive entertainment [3].

With the rapid development of deep learning, a variety of FER methods [2], [3] have been developed and achieved excellent classification performance in both in-the-lab and in-the-wild environments. Most of these methods focus on classi-

fying basic expression categories (including angry, disgusted, fearful, happy, sad, surprised, contempt, and neutral) [4].

Unfortunately, basic expressions fail to fully capture human emotions in real-life scenarios. To cover more human emotions, Du *et al.* [5] define compound expression categories (e.g., happily-surprised), where each compound expression is the combination of two basic expressions. Generally, compound expressions can describe human emotions more comprehensively. Unlike basic expressions, compound expressions show more subtle variations, and they are more challenging to be identified. Thus, most existing compound FER methods [6], [7] rely heavily on large-scale labeled compound expression data for training. However, annotating such data is time-consuming and labor-intensive due to subtle differences between compound expressions.

Recently, few-shot learning (FSL) has emerged as a promising learning scheme to avoid expensive annotations. Inspired by FSL, some methods [8], [9] study the cross-domain few-shot FER (CF-FER) task, which largely reduces the requirement of annotating large-scale compound expression data in conventional compound FER methods. They typically identify novel compound expressions (which involve only a limited number of reference images in the target domain) by using the model trained on multiple basic expression datasets in the source domain. By utilizing easily accessible basic expression datasets, such a task greatly alleviates the heavy burden of expensive annotation costs and expands potential applications of compound FER. In this paper, we study the CF-FER task following the same settings as the above methods [8], [9].

Generally, the hierarchical relationships between images are very common in many computer vision tasks [10], [11]. For CF-FER, there also exist rich hierarchical relationships of facial expressions. As illustrated in Fig. 1(a), a compound expression (e.g., happily-disgusted) can be described as a combination of two basic expressions (e.g., happy and disgusted). The happy and disgusted expressions belong to different emotional valences (i.e., the positive and negative expressions, respectively). In fact, humans can easily learn the hierarchical structure of facial expressions and apply it to identify new compound expressions with only a few images. Therefore, exploiting hierarchical relationships of expressions for model learning not only facilitates the full utilization of limited data in FSL, but also encourages the model to understand facial expressions comprehensively. This greatly improves the model's capability to transfer knowledge from basic expressions to novel compound expressions.

Existing CF-FER methods [8], [9] leverage the multi-dataset to increase the diversity of the source domain and learn a transferable space. Unfortunately, these methods often learn feature embeddings in the Euclidean space without considering imbalanced expression categories and imbalanced sample difficulty in the multi-dataset. As a result, they are prone to focus on learning the expression categories involving a larger number of samples and ignore the distinction between easy and hard samples, failing to sufficiently capture the inherent hierarchical relationships of facial expressions. This significantly reduces the model's transferability.

To address these issues, we propose a hyperbolic self-paced multi-expert network (HSM-Net) for CF-FER. HSM-Net consists of multiple mixture-of-experts (MoE) convolutional layers, where each expert focuses on learning a relatively balanced subset of expression categories from the multi-dataset. Such a way effectively reduces the issue of imbalanced expression categories. Based on it, we introduce a hyperbolic self-paced learning (HSL) strategy. The strategy projects features from the Euclidean space to the hyperbolic space and leverages the inherent geometric properties of hyperbolic space to capture the hierarchical structure of expressions. According to the Riemannian gradient update characteristics in the hyperbolic space, we can naturally perform self-paced learning to train the model from easy-to-hard samples, thereby alleviating the issue of imbalanced sample difficulty.

Under the above designs, we can learn effective feature embeddings in the hyperbolic space by reducing the influence of both imbalanced expression categories and imbalanced sample difficulty. As a result, the hierarchical relationships of facial expressions are fully exploited to improve the generalization performance of our model on the target domain. Fig. 1(b) illustrates the feature distributions in the Euclidean and hyperbolic spaces. The Euclidean space treats feature embeddings of different expression categories equally. In contrast, the hyperbolic space appropriately models hierarchical relationships of expressions, revealing the intrinsic connections between basic and compound expressions.

Specifically, each MoE layer consists of a parameter-shared router and a vanilla convolutional layer from the backbone CNN model. The router is composed of a preference score estimation network (PSE-Net) to estimate preference scores for all experts, and an expert selection network (ES-Net) to select channel features from the backbone network based on these preference scores. In this way, each expert is adaptively learned to focus on specific expression categories in the multi-dataset. To enable the router to select the optimal expert for each facial image in the training set, we propose a preference loss. By optimizing the preference loss across multiple MoE layers, these experts are collaboratively trained to capture feature representations from relatively balanced expression category subsets, effectively mitigating the influence of imbalanced expression categories. The HSL strategy trains the model in a self-paced manner, addressing the influence of imbalanced sample difficulty. Based on the above, our method jointly alleviates the issues of imbalanced expression categories and imbalanced sample difficulty by incorporating the geometric properties and optimization mechanisms of hyperbolic space into the multi-expert network. Such a manner can facilitate the modeling of hierarchical relationships among expressions and enable the learning of a highly transferable feature space.

In summary, our main contributions are given as follows:

- We propose a novel HSM-Net for FER under the cross-domain few-shot settings. Our HSM-Net can substantially reduce the influence of imbalanced expression categories and imbalanced sample difficulty in the multi-dataset, thereby effectively capturing the hierarchical relationships of facial expressions.

- We develop an HSL strategy to train the model adaptively from easy-to-hard samples in the hyperbolic space. Our HSL strategy greatly enhances the model's transferability from seen basic facial expressions to unseen compound facial expressions.
- We conduct extensive experiments on both in-the-lab and in-the-wild compound expression datasets to show the effectiveness of our HSM-Net over several state-of-the-art FSL methods for the CF-FER task.

The remainder of this paper is organized as follows. First, Section II briefly reviews the related work. Then, Section III presents the details of our proposed method. Next, Section IV provides extensive experimental results on compound expression datasets. Finally, Section V gives the conclusion.

II. Related Work

In this section, we briefly review compound FER, few-shot FER, imbalanced learning, self-paced learning, and hyperbolic deep learning, which are closely related to our method.

A. Compound FER

Following Ekman and Friesen's work [4], [12], the conventional FER task classifies an input facial image into one of the basic expression categories. We define such a task as the basic FER task. A variety of FER methods [2], [3] concentrate on the basic FER task. Regrettably, basic expressions cannot completely characterize the diversity of human emotions in nature since human emotions involve compound expressions. Du *et al.* [5] reveal that facial images contain compound expressions, which can provide a more subtle distinction between different human emotions. Compound expressions are highly useful for accurately capturing and understanding human emotional states in practical applications.

Compared with the basic FER task, the compound FER task aims to identify compound expressions containing subtle variations. Compound FER is still in its infancy, leaving room for improvement. Li *et al.* [6] collect a real-world affective facial database annotated with compound emotions (RAF-CE), where both compound expression labels and AU labels are provided. Moreover, they propose a meta-based multi-task learning (MML) method for the compound FER task. Jiang *et al.* [7] propose an expression soft label mining (ESLM) method to address the negative influence of hard expression labels. Dong *et al.* [13] design a bi-center loss, which encourages deep neural networks to learn compound emotion features.

The above methods usually rely on abundant annotated compound expression data for training. Unfortunately, annotating high-quality compound expression data is time-consuming and often requires guidance from experts in psychology. Unlike these methods, we study compound FER under the cross-domain FSL setting, which largely alleviates the heavy burden of acquiring large-scale compound expression training data.

B. Few-Shot FER

Few-shot FER aims to identify new expressions with an extremely limited number of samples. Ciubotaru *et al.* [14] first study the mainstream FSL methods on basic expression datasets, exploring the feasibility of few-shot FER. Shome *et al.* [15] propose a few-shot federated learning framework for FER under decentralized training. The above methods apply FSL to the classification of basic expressions. Later, Zou *et al.* [8] are the first to study the CF-FER task and propose a dual-branch emotion guided similarity network (EGS-Net) to perform knowledge transfer from the source domain to the target domain. Subsequently, Zou *et al.* [9] develop a cascaded decomposition network (CDNet) that cascades several learn-to-decompose modules with shared parameters to obtain a transferable feature space. Chen *et al.* [16] introduce a self-supervised visual Transformer (SSF-ViT) based on self-supervised learning (SSL) and FSL, enabling the training of models with fewer labeled samples.

Following the same settings as [8], [9], we study the CF-FER task. However, different from existing methods, we focus on addressing the challenge of capturing the inherent hierarchical relationships of facial expressions caused by imbalanced expression categories and imbalanced sample difficulty in the Euclidean space, aiming to improve the generalization of the model on the unseen target domain.

C. Imbalanced Learning and Self-Paced Learning

Facial expression datasets often involve significantly imbalanced expression categories. To address this, existing methods can be roughly divided into data pre-processing, re-weighting, and model ensemble methods. This paper mainly studies model ensemble methods. To tackle class imbalance and expression similarity in both source and target domains, Yang *et al.* [17] propose a residual attentive sharing network (RASN), which introduces a shared affinity feature module to compensate for inadequate feature learning of minority classes. Sreenivas *et al.* [18] propose a method to tackle class imbalance and expression similarity in both source and target domains. Unlike traditional methods that combine multiple models to improve the performance, we introduce a mixture-of-experts (MoE) layer, which enhances robustness and generalization by dynamically activating a subset of specialized networks for each input.

The idea of self-paced learning is to simulate the human learning process, which generally starts by learning simpler samples of a learning task and then gradually introduces more complex examples into training, effectively addressing imbalanced sample difficulty. Zhang *et al.* [19] propose a progressive learning strategy to extend the conventional one-stage meta learning into a multi-stage training process. Recently, self-paced learning is also introduced to FER. Shao *et al.* [20] develop a self-paced label distribution learning strategy, which initially focuses on learning easy samples with reliable label distributions and gradually progresses to more complex samples. This reduces the negative influence caused by noisy samples and unreliable label distributions. In contrast to traditional self-paced learning methods that are based on the Euclidean space, we study hyperbolic self-paced learning, which employs hyperbolic uncertainty to determine the algorithmic learning pace. Such a way is beneficial for accurately learning hierarchical relationships of facial expressions.

D. Hyperbolic Deep Learning

Although the Euclidean space has been the standard way to learn visual representations, its inherent properties are not suitable for all types of data. For hierarchical structures, hyperbolic geometry can provide a direct fit [11]. Recent studies have shown that the hyperbolic space has advantages in hierarchical representation learning. Accordingly, hyperbolic deep learning has made rapid progress across various vision tasks. Dai *et al.* [21] propose a hyperbolic-to-hyperbolic graph convolutional network (H2H-GCN), which operates directly on hyperbolic manifolds, avoiding the distortion caused by tangent space approximations and preserving the global hyperbolic structure. Li *et al.* [22] propose a simple yet effective method to capture the hierarchical relationships between images for the few-shot image generation task by using data from seen categories in the hyperbolic space.

The task of FSL is concerned with the generalization performance of the model to adapt to unseen data. For the few-shot FER task, existing methods [8], [9], [15] are often based on metric learning, which computes the Euclidean distance between image representations extracted by deep neural networks as a measure of similarity. On the contrary, we study few-shot FER in the hyperbolic space, which models the inherent hierarchical relationships of facial expressions in the source domain, thereby facilitating improving the model's transferability in the unseen target domain based on the close correlations between basic and compound expressions.

III. Method

In this section, we introduce our HSM-Net in detail. First, we give the problem formulation in Section III-A. Then, we provide an overview of HSM-Net in Section III-B. Next, we present the key components of HSM-Net in Sections III-C and III-D. Finally, we give the overall loss in Section III-E.

A. Problem Definition

In this paper, we consider the compound FER task in the cross-domain FSL setting, where only a few novel class samples are required to identify a compound expression category in the target domain. Given a labeled training set (the source domain) $\mathcal{D}_{train} = \{\mathcal{X}_{train}, \mathcal{Y}_{train}\}$, consisting of C_{base} base classes. Here, $\mathcal{X}_{train} = \{\mathbf{x}_i\}_{i=1}^{T}$ denotes the basic facial expression images, $\mathcal{Y}_{train} = \{\mathbf{y}_i^t\}_{i=1}^{T}$ is the set of ground-truth labels, and T represents the number of training samples. For the test set (the target domain), we denote it as $\mathcal{D}_{test} = \{\mathcal{X}_{test}, \mathcal{Y}_{test}\}$, consisting of C_{novel} novel classes.

In this paper, we aim to learn a model on the source domain so that it can be well generalized to the target domain. Following previous settings [8], [9], the base classes refer to the basic expression categories, while the novel classes refer to the compound expression categories. To enrich the diversity of the training set and alleviate the discrepancy between the source and target domains, multiple easily accessible basic expression datasets are used as the source domain. Note that the base classes and novel classes are disjoint and the number of base classes is limited in the source domain.

The few-shot task classifies the query images with the reference of the support images. After training the model on $\mathcal{D}_{train}$, we design multiple few-shot learning tasks on $\mathcal{D}_{test}$ to evaluate the performance of the model learned in the target domain. Each few-shot task (an N-way K-shot task) samples N classes from C_{novel} classes, and each class contains K labeled support samples and Q unlabeled query samples. Similar to representative FSL methods [36], we assign the query images to their nearest classes in the learned expression feature space.

B. Overview

In this paper, we develop a hyperbolic self-paced multi-expert network (HSM-Net), which contains multiple mixture-of-experts (MoE) convolutional layers located in the hyperbolic space, for CF-FER. HSM-Net collaboratively trains multiple experts in a self-distillation manner, where each expert focuses on classifying only a subset of expression categories. Such a way effectively addresses the influence of imbalanced expression categories. Based on it, we further introduce a hyperbolic self-paced learning (HSL) strategy to transform the feature space from the Euclidean space to the hyperbolic space and formulate the training process as self-paced learning. Thus, we can adaptively train the model from easy-to-hard samples, mitigating the influence of imbalanced sample difficulty.

Based on the above designs, our method can capture hierarchical relationships of facial expressions in the hyperbolic space by reducing the influence of imbalanced expression categories and imbalanced sample difficulty. As a result, the model's transferability from basic facial expressions to compound expressions is greatly enhanced and thus a highly transferable feature space is learned. The overview of our method is shown in Fig. 2.

C. Hyperbolic Self-Paced Multi-Expert Network (HSM-Net)

HSM-Net incorporates multiple MoE layers, where each layer includes a parameter-shared router and a vanilla convolutional module derived from the backbone CNN (we use ResNet [23] in this paper). The router includes a preference score estimation network (PSE-Net) to estimate preference scores for experts and an expert selection network (ES-Net) to select a subset of channel features from the backbone network based on preference scores. By enforcing each expert to focus on a relatively balanced subset of expression categories, the model can effectively learn appropriate decision boundaries to reduce the influence of imbalanced expression categories in the multi-dataset.

For each MoE layer, we define E experts, each of which is responsible for learning representations from a subset of expression categories. Suppose that the original backbone network has C_1 channels in the first convolutional layer. By scaling the model (i.e., varying the number of kernels) with the ratio r ($0 < r \leq 1$), we update the first convolutional layer containing rEC_1 channels, where each expert selects rC_1 channels in the first convolutional layer. Similarly, the i-th convolutional layer in the updated backbone network contains

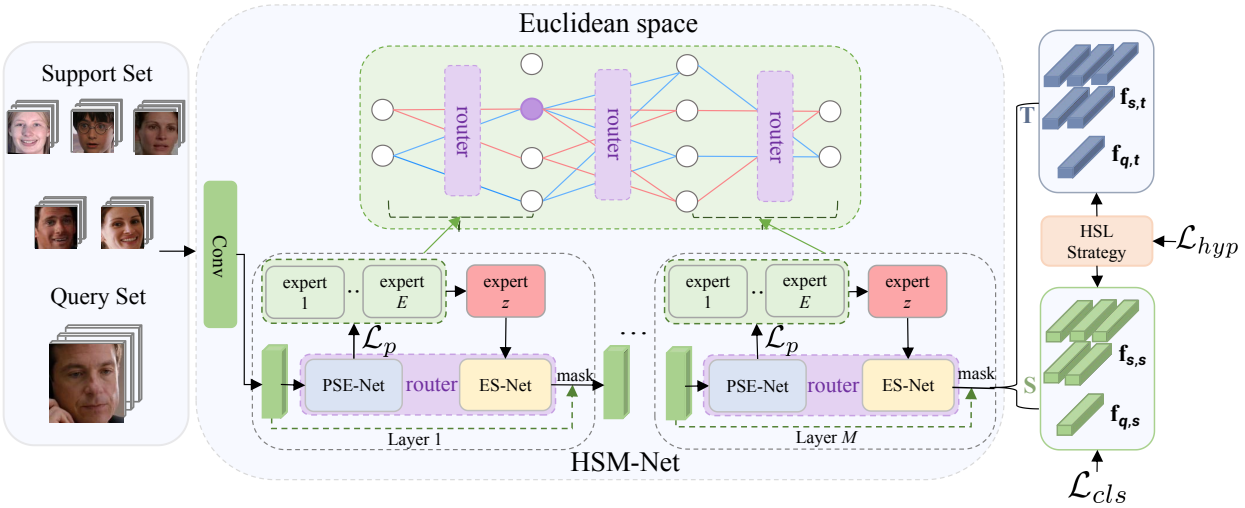


Fig. 2: Overview of our proposed HSM-Net. HSM-Net consists of multiplemixture-of-experts (MoE) convolutional layers, where each layer includes a parameter-shared router and a vanilla convolutional module derived from the backbone CNN. The router contains a preference score estimation network (PSE-Net) and an expert selection network (ES-Net) to estimate preference scores for experts and select a subset of channel features from the backbone network, respectively. In this figure, 'Conv' denotes the convolutional layer, **T** represents the teacher model and **S** represents the student model.

rEC_i channels, where C_i is the channel number in the i-th convolutional layer of the original backbone network.

Specifically, given an input image $\mathbf{x}_i \in \mathbb{R}^{3\times H\times W}$ in the training set, we first pass it through a preprocessing block (i.e., containing a 2×2 convolutional layer, followed by a ReLU activation function and an average pooling operation) of the backbone network to extract the basic feature $\mathbf{f}_i^0$. Then, $\mathbf{f}_i^0$ is fed into PSE-Net in the router which calculates the preference scores $\mathbf{s}_i^1 \in \mathbb{R}^E$ for E experts in the first convolutional layer of the backbone network through multiple parallel convolutional blocks (i.e., each block consists of a 2×2 convolutional layer, an average pooling layer, and a fully connected layer). Then, the expert z with the highest preference score is selected.

Next, ES-Net generates a mask $\mathbf{m}_i^1 \in \{0,1\}^{rEC_1}$, which is used to select the subset of channels corresponding to the selected expert z. The element in the mask is set to 1 when a channel is selected and 0 otherwise. Finally, we apply $\mathbf{m}_i^1$ to extract the channel feature (denoted as $\mathbf{f}_i^1$) for the expert z. $\mathbf{f}_i^1$ is served as the input feature for the second convolutional layer. Analogously, for each subsequent convolutional layer of the backbone network, the feature from the previous convolutional layer is fed into the router to select the optimal expert. These selected channels then form an end-to-end pathway, building an expert model for $\mathbf{x}_i$. The final expression feature extracted by this expert model is denoted as $\mathbf{f}_i^e$.

During the channel selection process, it is critical that the router can effectively give appropriate preference scores so that the experts can correctly focus on the subsets of expression categories. To achieve this, we leverage a cross-entropy loss based on the preference scores to constrain the experts selected by the router. Technically, inspired by hierarchical relationships of expressions, we first divide expression categories into three subsets (positive, neutral, and negative expression subsets) according to emotional valence. To balance the number of samples in each subset, we further split the expression subset into several smaller subsets to ensure the balanced number of samples in each subset. Each subset is learned by an expert. Hence, we define the loss as

$$\mathcal{L}_{ce} = \sum_{l=1}^{L} \mathrm{CE}(\mathbf{s}_i^l, y_e), \tag{1}$$

where $\mathrm{CE}(\cdot,\cdot)$ is the cross-entropy loss; L denotes the total number of convolutional layers; $\mathbf{s}_i^l$ represents the preference scores of the i-th input image calculated in the l-th convolutional layer; $y_e \in [1, E]$ denotes the index of the expert *w.r.t.* the input image $\mathbf{x}_i$.

To enhance the generalization performance of the expert model on compound expressions, we apply the flooding scheme [24] to the above loss, which intentionally prevents further reduction of the training loss when it reaches a reasonably small value. The preference loss is defined as

$$\mathcal{L}_p = |\mathcal{L}_{ce} - b| + b, \tag{2}$$

where $b > 0$ is the flood level used to control the range of loss fluctuation. With the flooding scheme, the model will continue to "random walk" with the same non-zero training loss, and drift into an area with a flat loss landscape. This allows the router to have some margins of errors when selecting experts (the router can choose the expert that does not correspond to the current sample category based on preference scores). Thus, each expert can learn not only fixed-category expressions but also information from other expression categories. This facilitates better generalization from basic expressions to unseen compound expressions.

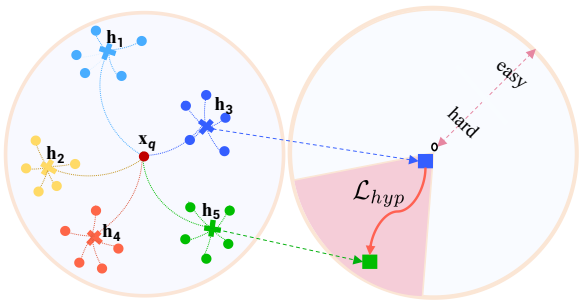


Fig. 3: Illustration of the Hyperbolic Self-Paced Learning (HSL) strategy. This strategy captures the hierarchical relationships of expression features in the hyperbolic space while training the model in a self-paced manner.

D. Hyperbolic Self-Paced Learning (HSL) Strategy

To stabilize the learning process, we adopt a self-distillation network architecture [25] to distill knowledge from the embedding model (denoted as the teacher model) into a new model with an identical architecture (denoted as the student model). Given an input image $\mathbf{x}_i$, we obtain expression features $\mathbf{f}^e_{i,t}$ and $\mathbf{f}^e_{i,s}$ from the teacher and student models, respectively. However, the feature embeddings learned by the teacher or student models are still in the Euclidean space and cannot capture hierarchical relationships of facial expressions. Moreover, the training process is further influenced by imbalanced sample difficulty, resulting in inferior transferability for the CF-FER task. Therefore, we introduce an HSL strategy to capture the hierarchical relationships of facial expression features in the hyperbolic space, where we can leverage sample difficulty to facilitate training from easy-to-hard samples, thus improving the transferability from basic expressions to compound expressions. The illustration of the HSL strategy is shown in Fig. 3.

Inspired by [26], we use the exponential map function $\mathrm{Exp}^c_o(\cdot)$ to project the features $\mathbf{f}^e_{i,t}$ and $\mathbf{f}^e_{i,s}$ from the Euclidean space to the hyperbolic embedding space (i.e., the Poincaré ball centered at the origin o), denoted as $\mathbf{h}_{i,t}$ and $\mathbf{h}_{i,s}$

$$\mathbf{h}_{i,t} = \mathrm{Exp}^c_o(\mathbf{f}^e_{i,t}) = \tanh(\sqrt{c}\|\mathbf{f}^e_{i,t}\|)\frac{\mathbf{f}^e_{i,t}}{\sqrt{c}\|\mathbf{f}^e_{i,t}\|}, \tag{3}$$

$$\mathbf{h}_{i,s} = \mathrm{Exp}^c_o(\mathbf{f}^e_{i,s}) = \tanh(\sqrt{c}\|\mathbf{f}^e_{i,s}\|)\frac{\mathbf{f}^e_{i,s}}{\sqrt{c}\|\mathbf{f}^e_{i,s}\|}, \tag{4}$$

where c represents the curvature of the hyperbolic space; $\|\cdot\|$ the standard Euclidean L_2-norm; $\tanh(\cdot)$ is the hyperbolic tangent function.

In the hyperbolic space, we aim to minimize the Poincaré distance between $\mathbf{h}_{i,t}$ and $\mathbf{h}_{i,s}$, that is,

$$\mathcal{L}_{hyp} = \cosh^{-1}\left(1 + 2\frac{\|\mathbf{h}_{i,s} - \mathbf{h}_{i,t}\|^2}{(1 - \|\mathbf{h}_{i,s}\|^2)\,(1 - \|\mathbf{h}_{i,t}\|^2)}\right), \tag{5}$$

where $\|\mathbf{h}_{i,s}\|$ and $\|\mathbf{h}_{i,t}\|$ respectively denote the radii of $\mathbf{h}_{i,s}$ and $\mathbf{h}_{i,t}$ in the Poincaré ball and $\cosh^{-1}(\cdot)$ denotes the inverse hyperbolic cosine function.

In the Poincaré ball, the local volume exponentially expands from the center to the boundary. This causes the learned feature embeddings to preferentially project hard samples near the center of the Poincaré ball, while moving easy samples towards the boundary of the Poincaré ball [10], [27]. In addition, the characteristic of the hyperbolic space allows us to use the center of the Poincaré ball as a reference point, where the hyperbolic uncertainty is defined as the distance from the embedding to the center. Therefore, the distance between an embedding and the center provides a natural estimation of sample difficulty. Based on the geometric characteristics of hyperbolic space, the input images with lower sample difficulty (i.e., easy samples) are mapped closer to the boundary, while images with higher sample difficulty (i.e., hard samples) are mapped closer to the center. This property aligns well with the nature of expression feature learning (where ambiguous, low-quality, or domain-shifted expression samples typically exhibit higher uncertainty), whereas typical and high-quality expression samples are more confidently classified. This representation of sample difficulty or uncertainty is termed hyperbolic uncertainty. Following [28], our method leverages hyperbolic uncertainty to indicate sample difficulty, facilitating training from easy-to-hard samples.

We update the teacher model using the exponential moving average of the student model. That is, $\phi = \omega\phi + (1-\omega)\phi'$, where ω is the controllable weight (we set it to 0.99); ϕ and ϕ' represent the network parameters of the teacher model and the student model, respectively. Thus, in the early training stage, the teacher model provides a more stable estimation of sample difficulty than the student model [29], [30]. In this way, we can define the sample difficulty of the hyperbolic embedding $\mathbf{h}_{i,t}$ *w.r.t.* the feature $\mathbf{f}^e_{i,t}$ as

$$u_{\mathbf{h}_{i,t}} = 1 - \|\mathbf{h}_{i,t}\|, \tag{6}$$

where $u_{\mathbf{h}_{i,t}}$ denotes the sample difficulty.

Optimization. To learn the model parameters, we employ stochastic Riemannian gradient descent [31] to minimize the Poincaré distance between $\mathbf{h}_{i,t}$ and $\mathbf{h}_{i,s}$. This is based on the Riemannian gradient of Eq. (5), which is computed *w.r.t.* the hyperbolic embedding $\mathbf{h}_{i,s}$ of the student model. The optimization procedure pushes the student embedding to match the teacher embedding $\mathbf{h}_{i,t}$

$$\nabla L_{hyp} = \frac{\left(1 - \|\mathbf{h}_{i,s}\|^2\right)^2}{2\sqrt{(1 - \|\mathbf{h}_{i,s}\|^2)\,(1 - \|\mathbf{h}_{i,t}\|^2) + \|\mathbf{h}_{i,s} - \mathbf{h}_{i,t}\|^2}} \times \left(\frac{\mathbf{h}_{i,s} - \mathbf{h}_{i,t}}{\|\mathbf{h}_{i,s} - \mathbf{h}_{i,t}\|} + \frac{\mathbf{h}_{i,s}\|\mathbf{h}_{i,s} - \mathbf{h}_{i,t}\|}{1 - \|\mathbf{h}_{i,s}\|^2}\right). \tag{7}$$

The above learning process is self-paced, where the gradient changes according to the sample difficulty $u_{\mathbf{h}_{i,t}}$ from the teacher model (Eq. (6)), i.e., the larger the radius $\|\mathbf{h}_{i,t}\|$ is, the easier $\mathbf{h}_{i,t}$ is, and the stronger the gradient ∇L_{hyp} is, regardless of $\mathbf{h}_{i,s}$. Such a way can achieve a training strategy from easy-to-hard samples.

Different from the Euclidean space, the hyperbolic space can naturally embed hierarchical structures [32], [33]. Therefore, during the optimization process of the hyperbolic space, the hierarchical relationships of facial expression features are implicitly captured by the multi-expert network (as illustrated

in Fig. 1(a)), encouraging the router to better understand facial features, calculate more accurate preference scores, and enhance the optimization of expert models. Therefore, learning high-level hierarchical structural information can enable the model to obtain an effective transferable space. Furthermore, without introducing additional cost as previous self-paced learning methods [34], [35], our HSL strategy adaptively assigns larger gradient changes to easy samples and smaller gradient changes to hard samples.

Finally, the expression feature obtained by the student model of our HSM-Net is used for expression classification. Same as ProtoNet [36], each query image is assigned to its nearest center of the support class in the learned feature space. The expression classification loss of a query image is

$$\mathcal{L}_{cls} = -\sum_{n=1}^{N} 1_{[n=y_q]} \log\left(\mathrm{softmax}\left(-\mathcal{M}\left(\mathbf{f}^e_{q,s}, \mathbf{R}_n\right)\right)\right), \quad (8)$$

where $\mathbf{f}^e_{q,s}$ and y_q are the final expression feature and the expression label of the query image, respectively; $\mathbf{R}_n = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{f}^e_{s,s})^n_k$ represents the center of class n and $(\mathbf{f}^e_{s,s})^n_k$ is the expression feature of the k-th image in class n in the support set; $\mathcal{M}(\cdot)$ denotes the metric module; N is the number of sampled classes; $\mathrm{softmax}(\cdot)$ denotes the softmax function.

E. Overall Loss

Based on the above, the overall loss is

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{cls} + \alpha h \mathcal{L}_{hyp} + \beta \mathcal{L}_p, \quad (9)$$

where α, h, and β are the balancing parameters. Note that the parameter α is dynamically set to $\lambda\frac{u}{U}$ (u represents the current epoch, U represents the total number of epochs, and λ denotes the trade-off parameter (we empirically set it to 0.5)). The choice of α is inspired by curriculum learning so that the model can initially focus on the classification loss and gradually shift to hyperbolic space optimization.

IV. Experiments

In this section, we first introduce the experimental settings, including facial expression datasets and implementation details in Section IV-A. Then, we perform ablation studies and give some visualization results in Section IV-B. Finally, we compare our method with several state-of-the-art methods and discuss the limitations of our method in Section IV-C.

A. Experimental Settings

Datasets. Our HSM-Net is trained on multiple basic expression datasets and tested on the compound expression dataset. To ensure the diversity of basic expressions, we use five basic expression datasets, including three in-the-lab datasets: CK+ [37], MMI [38], and Oulu-CASIA [39], as well as two in-the-wild datasets: RAF-DB [40] and SFEW [41] to construct the training set. Three compound expression datasets (CFEE_C [5], EmotioNet_C [42], and RAF_C [40]) are used to evaluate the performance of the learned model.

Basic Expression Datasets. CK+ contains 593 video sequences from a total of 123 different subjects, where 327 video sequences are annotated with seven basic expressions. MMI contains 326 video sequences with six basic expressions (205 frontal-view sequences are used). Oulu-CASIA consists of 2,880 video sequences with six basic expressions (480 normal indoor illumination sequences are used). Three peak frames of each sequence in the above in-the-lab datasets are selected. RAF-DB consists of a basic subset with seven basic expressions and a compound subset with 11 compound expressions. The basic subset contains 12,271 training images. SFEW is labeled with seven basic expressions with 958 training images. All the samples in these basic expression datasets are used for training. Note that the imbalance ratio of the training set is up to 1:114 (the ratio between the contempt expressions and the happy expressions), indicating significantly imbalanced expression categories.

Compound Expression Datasets. CFEE_C is derived from the CFEE dataset. It is an in-the-lab dataset and annotated with 15 compound expressions for 230 subjects, including a total of 5,046 facial expression images. EmotioNet_C is collected from the EmotioNet challenge, where the samples (including 2,471 facial expression images) are collected in the wild and annotated with ten compound expressions. RAF_C is the compound subset of RAF-DB with 11 compound expressions and a total of 3,162 facial expression images.

Implementation Details. Our method is implemented by PyTorch. All the facial images in the training set are first aligned and resized to the size of 256×256. Then, they are randomly cropped to the size of 224×224, followed by a random horizontal flip and color jitter as data augmentation. Following [8], [9], we use a mapping function to unify the expression labels in basic expression datasets. In the training set, the number of images in the negative expression subset is larger than those in the positive and neutral subsets. Hence, we split the negative expression subset into two smaller subsets to ensure the balanced expression categories. Therefore, the number of experts E is set to 4. We do not employ pre-trained models in our method. We use the identical network structures for both the teacher model and the student model for self-distillation, where we update the teacher model by momentum updates.

We employ RFS [25] as our baseline method, where ResNet-12 is used as the backbone. Our model is trained using the stochastic gradient descent (SGD) optimizer with a learning rate of 0.1, and the weight decay is set to 5×10^{-4}. For the training stage, the model is optimized by 100 epochs of self-distillation batch training. For a few-shot task, following RFS [25], we only update the parameters of the predictor based on the support set and evaluate the model on the query set while keeping the backbone parameters fixed. We set the number of classes N = 5, the number of support samples K = 1 or 5, and the number of query samples Q = 16 for each class. The flood level b in Eq. (2) is set to 0.7. The curvature of the hyperbolic space c in Eq. (3) and Eq. (4) is set to 1.0. The balancing parameters α, h and β in Eq. (9) are set to 0.5, 0.01 and 0.5, respectively. The ratio r is set to 0.3.

TABLE I: Details of six variants of our HSM-Net and the corresponding ablation study results on the CFEE_C and RAF_C datasets. The test accuracy (%) of 5-way few-shot classification tasks with 95% confidence intervals is reported.

Methods	Details of Variants						CFEE_C		RAF_C	
	Baseline	MoE	Router	$\mathcal{L}_p$	HSL	$\mathcal{L}_{hyp}$	1-shot	5-shot	1-shot	5-shot
Baseline	√	×	×	×	×	×	54.96 ± 0.73	65.71 ± 0.61	43.05 ± 0.59	60.08 ± 0.46
HSM-Net_w/o_$_{(R+H)}$	√	√	×	×	×	×	55.74 ± 0.86	67.21 ± 0.70	45.18 ± 0.60	61.82 ± 0.43
HSM-Net_w_$_R$	√	√	√	×	×	×	56.35 ± 0.87	68.03 ± 0.71	46.07 ± 0.61	62.68 ± 0.45
HSM-Net_w_$_{\mathcal{L}_p}$	√	√	√	√	×	×	57.09 ± 0.87	68.92 ± 0.72	46.90 ± 0.62	63.49 ± 0.46
HSM-Net_w_$_H$	√	√	×	×	√	√	57.17 ± 0.86	68.97 ± 0.70	46.96 ± 0.62	63.44 ± 0.43
HSM-Net	√	√	√	√	√	√	57.96 ± 0.86	69.89 ± 0.70	48.02 ± 0.60	64.23 ± 0.45

B. Ablation Studies

We evaluate the performance obtained by six variants of our proposed method, including: 1) the baseline method; 2) the method (denoted as HSM-Net_w/o_$_{(R+H)}$) that only constructs multiple experts to classify samples without using routers and the HSL strategy; 3) the method (denoted as HSM-Net_w_$_R$) that employs routers to randomly select experts without the preference loss and the HSL strategy; 4) the method (denoted as HSM-Net_w_$_{\mathcal{L}_p}$) that employs the routers and the preference loss $\mathcal{L}_p$ in Eq. (1); 5) the method (denoted as HSM-Net_w_$_H$) that employs the HSL strategy and $\mathcal{L}_{hyp}$ to capture hierarchical relationships of facial expressions and train the model from easy-to-hard; 6) our proposed method (denoted as HSM-Net) that adopts the routers and the HSL strategy by optimizing both $\mathcal{L}_p$ and $\mathcal{L}_{hyp}$. The details of six variants of our method and the corresponding ablation study results are summarized in Table I. We use the CFEE_C and RAF_C datasets for ablation studies.

Effectiveness of the Mixture-of-Experts (MoE) Convolutional Layers. As observed from Table I, the HSM-Net_w/o_$_{(R+H)}$ method obtains better performance than the baseline method on the CFEE_C and RAF_C datasets. Specifically, compared with the baseline method, the HSM-Net_w/o_$_{(R+H)}$ method improves the performance by 0.78% on CFEE_C and 2.13% on RAF_C for the 5-way 1-shot classification task. These results show the effectiveness of learning multiple MoE layers that can focus on different expression categories, reducing the influence of imbalanced expression categories and enhancing the final performance.

Effectiveness of the Router. Compared with the HSM-Net_w/o_$_{(R+H)}$ method, HSM-Net_w_$_R$ improves the accuracy by 0.61% and 0.89% on CFEE_C and RAF_C, respectively, for the 5-way 1-shot classification task. The router can enable the model to learn accurate preference scores, encouraging the experts to focus on the expression categories they excel at handling. Therefore, a more appropriate decision boundary can be obtained, and the generalization performance on the target domain is improved. In addition, as shown in Table I, the HSM-Net_w_$_{\mathcal{L}_p}$ method gives better results than the HSM-Net_w_$_R$ method, showing that selecting the optimal expert for input images with the preference loss contributes to the final performance. This validates the effectiveness of the preference loss.

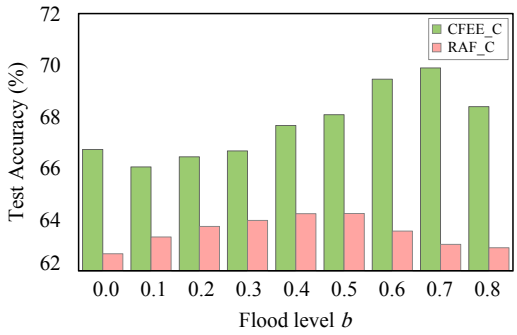
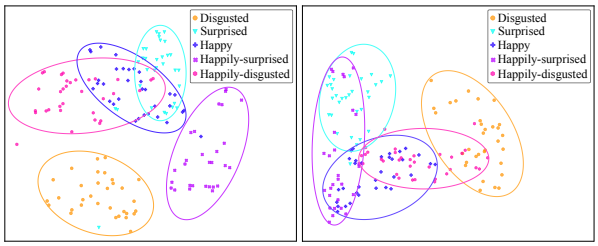


Fig. 4: Influence of the flood level b for the 5-way 5-shot classification task on the CFEE_C and RAF_C datasets. The best results are obtained when the value of b is set to 0.7 on the CFEE_C dataset and 0.5 on the RAF_C dataset.

TABLE II: Ablation study results of different data-splitting strategies on the CFEE_C and RAF_C datasets. Test accuracy (%) of 5-way few-shot classification tasks with 95% confidence intervals is reported.

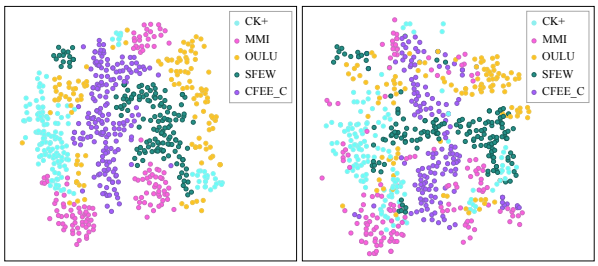
Methods	CFEE_C		RAF_C	
	1-shot	5-shot	1-shot	5-shot
Strategy 0	55.21 ± 0.72	64.65 ± 0.63	44.03 ± 0.61	62.22 ± 0.42
Strategy 1	56.44 ± 0.70	66.11 ± 0.62	46.12 ± 0.61	62.67 ± 0.44
Strategy 2	55.66 ± 0.70	65.27 ± 0.62	44.30 ± 0.60	62.52 ± 0.43
Our Strategy	57.96 ± 0.86	69.89 ± 0.70	48.02 ± 0.62	64.23 ± 0.42

Effectiveness of the Hyperbolic Self-Paced Learning (HSL) Strategy. Compared with the HSM-Net_w/o_$_{(R+H)}$ method, HSM-Net_w_$_H$ improves the accuracy by 1.43% and 1.78% on CFEE_C and RAF_C, respectively, for the 5-way 1-shot classification task. This indicates that our HSL strategy can effectively faciliate the model to capture hierarchical relationships of facial expressions in the hyperbolic space. Such hierarchical relationships provide useful structural information, enabling the network to better generalize to new compound expressions. Moreover, the HSL strategy further enhances the model's generalization performance by training the model from easy-to-hard samples. Meanwhile, as shown



(a) Baseline (b) HSM-Net

Fig. 5: Visualization of the features of different expression categories extracted by the baseline method and our HSM-Net method on the CFEE_C dataset under the 5-way 5-shot setting. Different colors represent different facial expression categories.



(a) Baseline (b) HSM-Net

Fig. 6: Visualization of the extracted features obtained by (a) the baseline method and (b) HSM-Net on multi-source domains and the target domain (CFEE_C), with different colors representing different datasets. Our HSM-Net reduces the domain discrepancy between the multiple source domains and the target domain, thereby enhancing the cross-domain FER task.

in Table I, the HSM-Net method achieves better results than HSM-Net_w_$\mathcal{L}_p$ methods. The above results demonstrate the effectiveness of the HSL strategy.

Influence of Different Data-Splitting Strategies. We evaluate the influence of different data-splitting strategies on the final performance. The results are given in Table II. Strategy 0 refers to the strategy that divides the dataset into four subsets using a random data-splitting method without considering emotional valence or sample size. Strategy 1 refers to the strategy that divides the dataset into three subsets (positive, negative, and neutral expression subsets) based on emotional valence. Strategy 2 refers to the strategy that divides the datasets into 4 subsets to ensure a balanced sample size according to the number of class samples instead of emotional valence. Our strategy refers to the proposed strategy that divides the datasets into 4 subsets by considering both emotional valence and sample size.

We can see that Strategy 0 yields the worst model performance. In contrast, Strategy 1 and Strategy 2 significantly improve the recognition accuracy. Note that Strategy 1 models the hierarchy of facial expressions while Strategy 2 mitigates the adverse influence of class imbalance during model training. Our strategy not only captures the hierarchical structure of facial expressions but also addresses class imbalance, achieving the best model performance across all the tasks. Our strategy provides a clearer semantic structure by exploiting prior expression knowledge and a more stable data distribution, enhancing the model's robustness and generalization ability.

Influence of the Flood Level b. We evaluate the influence of different values of the flood level b for the 5-way 5-shot classification task on the CFEE_C dataset. The performance obtained by our HSM-Net method with the different values of b is given in Fig. 4. The best results are obtained when the value of b is set to 0.7 on the CFEE_C dataset and 0.5 on the RAF_C dataset.

Influence of the Sample Size. We evaluate the influence of the sample size on the CFEE_C, EmotioNet_C, and RAF_C datasets. The results are given in Table III, where we report the test accuracy for 5-way 1-shot, 5-way 5-shot, and 5-way 10-shot classification tasks.

As shown in Table III, the classification performance is significantly boosted as the number of training samples per class is increased from 1 to 10. For example, the accuracy is improved from 48.02% (1-shot) to 68.00% (10-shot) on the RAF_C dataset. A larger sample size often provides more expression information, allowing the model to learn discriminative features and generalize to novel compound expression recognition tasks more effectively.

Influence of the Number of Layers. We evaluate the influence of the number of layers in the router network on model performance. We test the model performance with the different numbers of layers (including 1, 2, 3, and 4) for the 5-way 1-shot classification task on the CFEE_C dataset. The performance obtained by our HSM-Net method with the different numbers of layers d is given in Fig. 7.

We can see that the number of layers in the router network significantly affects the model performance. When the router network contains only 1 or 2 layers, its ability to learn complex routing decisions is limited, leading to relatively low performance. When the number of layers is set to 3, the model achieves the best accuracy, indicating that a deeper routing network enhances feature selection and expert assignment. However, when the number of layers is set to 4, the performance slightly drops. This is because of the increased model complexity, leading to overfitting under the few-shot setting. These results suggest that the appropriate number of layers can balance complexity and effectiveness in routing expert selection.

Influence of the Hyperparameter c. We evaluate the influence of the hyperparameter c in the hyperbolic space on model performance. We test the model performance with the different values of c (including 0.00001, 0.1, 0.5, 1, 5, and 10) for the 5-way 1-shot classification task on the CFEE_C dataset. The results are given in Table IV.

As shown in Table IV, the choice of c significantly affects model performance. When the value of c is too small (e.g., c = 0.00001), the hyperbolic space becomes nearly the Euclidean space. This limits the model's ability to capture hierarchical relationships and results in suboptimal performance. As the value of c is larger, the model benefits from hyperbolic

TABLE III: Ablation study results on the influence of the sample size on the CFEE_C, EmotioNet_C, and RAF_C datasets. Test accuracy (%) of 5-way 1-shot, 5-way 5-shot, and 5-way 10-shot classification tasks with 95% confidence intervals is reported.

Method	CFEE_C			EmotioNet_C			RAF_C		
	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
HSM-Net	57.96 ± 0.86	69.89 ± 0.70	74.12 ± 0.67	57.33 ± 0.66	64.95 ± 0.58	66.59 ± 0.67	48.02 ± 0.60	64.23 ± 0.45	68.00 ± 0.54

TABLE IV: Ablation study results of different hyperbolic space hyperparameter c on the CFEE_C dataset. Test accuracy (%) of 5-way 1-shot classification task with 95% confidence intervals is reported.

c	0.00001	0.1	0.5	1	5	10
1-shot	57.52 ± 0.82	57.64 ± 0.85	57.72 ± 0.85	57.96 ± 0.86	57.64 ± 0.86	52.94 ± 0.85

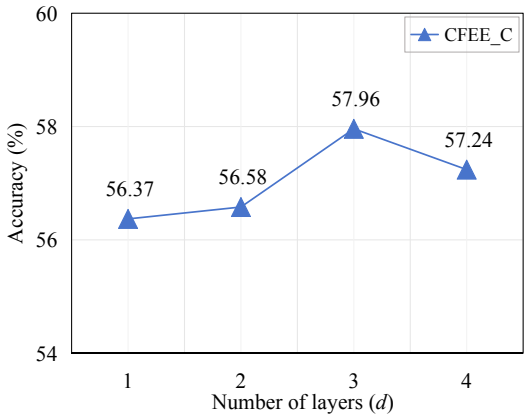


Fig. 7: Influence of the number of layers (d) in the router network for the 5-way 1-shot classification task on the CFEE_C dataset. The best results are obtained when the number of layers d is set to 3.

representations, achieving the highest accuracy at $c = 1$. However, when the value of c becomes too large (e.g., $c = 10$), the performance degrades significantly. This is because the excessive curvature value can distort feature representations and hinder effective classification. These observations suggest that a moderate curvature value offers the optimal trade-off between feature expressiveness and model stability in the hyperbolic space.

Visualization of the Hierarchical Relationships of Facial Expressions. We visualize the features of different expression categories extracted by the baseline method and HSM-Net on the CFEE_C dataset under the 5-way 5-shot setting, as shown in Fig. 5. As shown in Fig. 5(a), the features of different expression categories obtained by the baseline method are more separated, especially between the "happily-disgusted" "happy" and "disgusted" facial expressions. This shows that the baseline method fails to capture the hierarchical structure of expressions. In contrast, in Fig. 5(b), the overlapped pink, blue, and yellow circles in HSM-Net indicate the strong correlations between "Happy", "Disgusted" and "Happily-disgusted". This shows that HSM-Net effectively models the hierarchical relationships of facial expressions in the hyperbolic space.

Visualization of the Learned Features by the Baseline Method and our HSM-Net. To visually demonstrate the effectiveness of our HSM-Net, we visualize the features extracted by the baseline method and HSM-Net in the source domain and the target domain (CFEE_C) in Fig. 6. As shown in the figure, for the baseline method, the distribution gap between different datasets is distinct. In contrast, for HSM-Net, the distributions of different datasets (from both the source and target domains) are indistinguishable, indicating that the domain discrepancy is minimized. By collaboratively training multiple experts across different datasets, our method effectively addresses the problem of biased learning caused by imbalanced expression categories. In addition, the HSL strategy fully exploits the hierarchical relationships of facial expressions in the hyperbolic space based on a self-paced learning method, greatly improving the transferability from basic facial expressions to compound facial expressions.

C. Comparison with State-of-the-Art Methods

We compare our proposed HSM-Net with several state-of-the-art FSL methods, including episodic training-based FSL methods, batch training-based FSL methods, and hybrid FSL methods. For a fair comparison, our reported results of these competing methods are obtained by using the source codes provided by the respective authors under the same settings as ours, as done in Zou *et al*. [9].

From Table V, we can see that our HSM-Net method achieves the best results on the three datasets for both the 5-way 1-shot and 5-way 5-shot classification tasks. These results clearly validate the excellent performance of our method. Specifically, HSM-Net achieves an accuracy of 57.96% on CFEE_C, 57.33% on EmotioNet_C, and 48.02% on RAF_C in 5-way 1-shot classification tasks, and 69.89% on CFEE_C, 64.95% on EmotioNet_C, and 64.23% on RAF_C in 5-way 5-shot classification tasks, respectively. For batch training-based FSL methods, our method outperforms CDNet_B (which also employs the batch training strategy as ours) with improvements of 3.41%, 4.57%, and 6.00% on three datasets in 5-way 1-shot classification tasks, respectively. It is worth noting that our method achieves better performance than all hybrid FSL methods. This indicates that HSM-Net can obtain appropriate decision boundaries by using multiple experts, addressing the influence of imbalanced expression categories across multiple

TABLE V: Comparisons with state-of-the-art FSL methods on three different compound expression datasets. Test accuracy (%) of 5-way few-shot classification tasks with 95% confidence intervals is reported. The best and second-best results are marked in bold and underlined, respectively.

Methods	CFEE_C		EmotioNet_C		RAF_C	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
(a) Episodic training-based FSL methods						
ProtoNet [36]	53.29 ± 0.73	66.60 ± 0.60	50.15 ± 0.66	60.04 ± 0.56	39.12 ± 0.56	58.41 ± 0.46
MatchingNet [45]	52.31 ± 0.69	62.24 ± 0.61	48.64 ± 0.63	54.19 ± 0.58	34.84 ± 0.54	52.45 ± 0.44
RelationNet [46]	50.58 ± 0.68	63.17 ± 0.60	48.33 ± 0.68	56.27 ± 0.58	36.18 ± 0.54	53.45 ± 0.46
GNN [47]	54.01 ± 0.74	64.26 ± 0.63	49.49 ± 0.68	58.67 ± 0.59	38.74 ± 0.56	57.15 ± 0.47
DSN [48]	49.61 ± 0.73	60.03 ± 0.62	48.25 ± 0.68	54.89 ± 0.58	40.09 ± 0.55	52.49 ± 0.47
InfoPatch [49]	54.19 ± 0.67	67.29 ± 0.56	48.14 ± 0.61	59.84 ± 0.55	41.02 ± 0.52	57.98 ± 0.45
(b) Batch training-based FSL methods						
Softmax [50]	54.32 ± 0.73	66.35 ± 0.62	51.60 ± 0.68	61.83 ± 0.59	42.16 ± 0.59	58.57 ± 0.45
Cosmax [50]	54.97 ± 0.71	67.89 ± 0.61	50.87 ± 0.65	61.10 ± 0.56	40.87 ± 0.56	57.67 ± 0.46
Arcmax [51]	55.29 ± 0.71	67.72 ± 0.60	50.73 ± 0.65	61.70 ± 0.56	41.28 ± 0.57	57.94 ± 0.46
RFS [25]	54.96 ± 0.73	65.71 ± 0.61	51.91 ± 0.67	61.94 ± 0.57	43.05 ± 0.59	60.08 ± 0.46
LR+DC [52]	53.20 ± 0.73	64.18 ± 0.66	52.09 ± 0.70	60.12 ± 0.58	42.90 ± 0.60	56.74 ± 0.46
STARTUP [53]	54.89 ± 0.72	67.79 ± 0.61	52.61 ± 0.69	61.95 ± 0.57	43.97 ± 0.60	59.14 ± 0.47
CDNet_B [9]	54.55 ± 0.71	68.09 ± 0.62	52.76 ± 0.67	61.76 ± 0.57	42.02 ± 0.58	61.75 ± 0.44
(c) Hybrid FSL methods						
Meta-Baseline [54]	55.17 ± 0.74	67.15 ± 0.61	52.36 ± 0.67	62.01 ± 0.59	43.54 ± 0.61	61.59 ± 0.44
OAT [55]	54.28 ± 0.75	67.88 ± 0.62	52.92 ± 0.66	61.85 ± 0.59	42.75 ± 0.60	60.41 ± 0.43
BML [56]	52.42 ± 0.71	66.72 ± 0.61	51.31 ± 0.66	58.77 ± 0.57	41.91 ± 0.55	59.72 ± 0.45
EGS-Net [8]	56.65 ± 0.73	68.38 ± 0.60	51.62 ± 0.66	60.52 ± 0.56	44.07 ± 0.60	61.90 ± 0.46
CDNet [9]	<u>56.99</u> ± 0.73	<u>68.98</u> ± 0.60	<u>55.16</u> ± 0.67	<u>63.03</u> ± 0.59	<u>46.07</u> ± 0.59	<u>63.03</u> ± 0.45
HSM-Net (Ours)	**57.96** ± 0.86	**69.89** ± 0.70	**57.33** ± 0.66	**64.95** ± 0.58	**48.02** ± 0.60	**64.23** ± 0.45

TABLE VI: Comparison of accuracy (%) and inference time (s) obtained by several representative methods for the 5-way 5-shot classification task on the CFEE_C dataset.

Methods	Accuracy	Time
RFS	65.71	89.12
CDNet	68.98	95.58
HSM-Net	69.89	106.92

basic expression datasets. Meanwhile, it leverages sample uncertainty to guide training from easy-to-hard samples in the hyperbolic space, reducing the influence of imbalanced sample difficulty in the Euclidean space. As a result, HSM-Net enables the model to learn hierarchical relationships of facial expressions in the hyperbolic space and enhances its ability to generalize from seen basic expressions to unseen compound expressions.

Limitations. The comparison of accuracy and inference time obtained by several methods is given in Table VI. Although our HSM-Net achieves the best results in all few-shot tasks across the CFEE_C, EmotioNet_C, and RAF_C datasets, the introduction of the MoE layers increases computational load and inference time. The increased training time and resource demands may limit the scalability of our method, especially in resource-constrained environments. To address this issue, we adopt ResNet-12 as the backbone network instead of the typically higher-performing ResNet-18 [23]. While the selection of ResNet-12 reduces complexity, it also compromises the model's performance potential. Future work could explore further optimization of the MoE layers or methods to reduce model complexity without sacrificing performance. Potential directions include: (1) Dynamic expert sparsification that uses adaptive gating mechanisms to selectively activate only the most relevant experts for each sample [57], [58]. (2) Expert pruning that leverages pruning techniques to remove less influential experts or connections based on their contributions to the final performance [59], [60]. (3) Lightweight expert architectures that design efficient expert structures to reduce computational complexity while preserving model performance [61].

V. Conclusions

In this paper, we develop a novel HSM-Net for CF-FER by collaboratively training multiple experts across different datasets, effectively addressing the challenge of hierarchical relationship learning caused by imbalanced expression categories and imbalanced sample difficulties in the traditional Euclidean space. Based on the MoE layers, we introduce an HSL strategy to project features from the Euclidean space into the hyperbolic space, where we perform self-paced learning to train the model from easy-to-hard. As a result, our method can fully exploit hierarchical relationships of facial expressions and thus learn a transferable feature space. Experimental results on various compound FER datasets show the superiority of our method over several state-of-the-art FSL methods, validating the potential of learning transferable features in basic expression datasets for compound FER.

Currently, we leverage the simple ResNet-12 as the backbone. In future work, we will investigate more complicated backbones (such as Transformer) to investigate the applications of MoE for CF-FER.

REFERENCES

[1] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2006.

[2] D. Ruan, R. Mo, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Adaptive deep disturbance-disentangled learning for facial expression recognition," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 455–477, 2022.

[3] Y. Tang, X. Zhang, X. Hu, S. Wang, and H. Wang, "Facial expression recognition using frequency neural network," *IEEE Transactions on Image Processing*, vol. 30, pp. 444–457, 2020.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[5] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[6] X. Li, W. Deng, S. Li, and Y. Li, "Compound expression recognition in-the-wild with au-assisted meta multi-task learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5734–5743.

[7] J. Jiang, M. Wang, B. Xiao, J. Hu, and W. Deng, "Joint recognition of basic and compound facial expressions by mining latent soft labels," *Pattern Recognition*, vol. 148, pp. 110–173, 2024.

[8] X. Zou, Y. Yan, J.-H. Xue, S. Chen, and H. Wang, "When facial expression recognition meets few-shot learning: A joint and alternate learning framework," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 5367–5375.

[9] X. Zou, Y. Yan, J.-H. Xue, S. Chen, and H. Wang, "Learn-to-decompose: Cascaded decomposition network for cross-domain few-shot facial expression recognition," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 683–700.

[10] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, and V. Lempitsky, "Hyperbolic image embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6418–6428.

[11] R. Wei, Y. Liu, J. Song, Y. Xie, and K. Zhou, "Exploring hierarchical information in hyperbolic space for self-supervised image hashing," *IEEE Transactions on Image Processing*, vol. 33, pp. 1768–1781, 2024.

[12] P. Ekman and K. G. Heider, "The universality of a contempt expression: A replication," *Motivation and Emotion*, vol. 12, no. 3, pp. 303–308, 1988.

[13] R. Dong and K.-M. Lam, "Bi-center loss for compound facial expression recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 641–645, 2024.

[14] A.-N. Ciubotaru, A. Devos, B. Bozorgtabar, J.-P. Thiran, and M. Gabrani, "Revisiting few-shot learning for facial expression recognition," *arXiv preprint arXiv:1912.02751*, 2019.

[15] D. Shome and T. Kar, "Fedaffect: Few-shot federated learning for facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4168–4175.

[16] X. Chen, X. Zheng, K. Sun, W. Liu, and Y. Zhang, "Self-supervised vision transformer-based few-shot learning for facial expression recognition," *Information Sciences*, vol. 634, pp. 206–226, 2023.

[17] J. Yang, Z. Lv, K. Kuang, S. Yang, L. Xiao, and Q. Tang, "Rasn: using attention and sharing affinity features to address sample imbalance in facial expression recognition," *IEEE Access*, vol. 10, pp. 103 264–103 274, 2022.

[18] M. Sreenivas, S. Takamuku, S. Biswas, A. Chepuri, B. Vengatesan, and N. Natori, "Improved cross-dataset facial expression recognition by handling data imbalance and feature confusion," in *European Conference on Computer Vision*. Springer, 2022, pp. 262–277.

[19] L. Zhang, Z. Liu, W. Zhang, and D. Zhang, "Style uncertainty based self-paced meta learning for generalizable person re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 2107–2119, 2023.

[20] J. Shao, Z. Wu, Y. Luo, S. Huang, X. Pu, and Y. Ren, "Self-paced label distribution learning for in-the-wild facial expression recognition," in *Proceedings of the ACM International Conference on Multimedia*, 2022, pp. 161–169.

[21] J. Dai, Y. Wu, Z. Gao, and Y. Jia, "A hyperbolic-to-hyperbolic graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 154–163.

[22] L. Li, Y. Zhang, and S. Wang, "The euclidean space is evil: Hyperbolic attribute editing for few-shot image generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 714–22 724.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, "Do we need zero training loss after achieving zero training error?" *Proceedings of the International Conference on Machine Learning*, 2020, pp. 4604–4614.

[25] 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 *Proceedings of the European Conference on Computer Vision*, 2020, pp. 266–282.

[26] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," *Advances in Neural Information Processing Systems*, vol. 31, pp. 1–11, 2018.

[27] Y. Guo, X. Wang, Y. Chen, and S. X. Yu, "Clipped hyperbolic classifiers are super-hyperbolic classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11–20.

[28] D. Surís, R. Liu, and C. Vondrick, "Learning the predictability of the future," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 607–12 617.

[29] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[31] S. Bonnabel, "Stochastic gradient descent on riemannian manifolds," *Transactions on Automatic Control*, vol. 58, no. 9, pp. 2217–2229, 2013.

[32] F. Sala, C. De Sa, A. Gu, and C. Ré, "Representation tradeoffs for hyperbolic embeddings," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2018, pp. 4460–4469.

[33] R. Sarkar, "Low distortion delaunay embedding of trees in hyperbolic plane," in *Proceedings of the International Symposium on Graph Drawing*, 2011, pp. 355–366.

[34] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 7–19, 2017.

[35] K. Liu, W. Zhu, Y. Shen, S. Liu, N. Razavian, K. J. Geras, and C. Fernandez-Granda, "Multiple instance learning via iterative self-paced supervised contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3355–3365.

[36] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[37] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.

[38] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2005, pp. 317–321.

[39] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[40] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2852–2861.

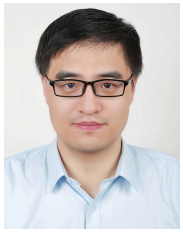
[41] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*. IEEE, 2011, pp. 2106–2112.

[42] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570.

[43] Y. Guo, H. Guo, and S. X. Yu, "Co-sne: Dimensionality reduction and visualization for hyperbolic data," in *Proceedings of the IEEE/CVF*

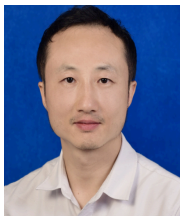
Conference on Computer Vision and Pattern Recognition, 2022, pp. 21–30.
[44] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.
[45] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3630–3638.
[46] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.
[47] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," in *Proceedings of the International Conference on Learning Representations*, 2018.
[48] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4136–4145.
[49] C. Liu, Y. Fu, C. Xu, S. Yang, J. Li, C. Wang, and L. Zhang, "Learning a few-shot embedding model with contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 8635–8643.
[50] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proceedings of the International Conference on Learning Representations*, 2019.
[51] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Associative alignment for few-shot image classification," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 18–35.
[52] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–13.
[53] C. P. Phoo and B. Hariharan, "Self-training for few-shot transfer across extreme task differences," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–19.
[54] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-Baseline: Exploring simple meta-learning for few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9062–9071.
[55] M. Chen, Y. Fang, X. Wang, H. Luo, Y. Geng, X. Zhang, C. Huang, W. Liu, and B. Wang, "Diversity transfer network for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 10 559–10 566.
[56] Z. Zhou, X. Qiu, J. Xie, J. Wu, and C. Zhang, "Binocular mutual learning for improving few-shot classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8402–8411.
[57] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.
[58] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "GShard: Scaling giant models with conditional computation and automatic sharding," *arXiv preprint arXiv:2006.16668*, 2020.
[59] X. Lu, Q. Liu, Y. Xu, A. Zhou, S. Huang, B. Zhang, J. Yan, and H. Li, "Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models," *arXiv preprint arXiv:2402.14800*, 2024.
[60] Y. Xie, Z. Zhang, D. Zhou, C. Xie, Z. Song, X. Liu, Y. Wang, X. Lin, and A. Xu, "MoE-Pruner: Pruning mixture-of-experts large language model using the hints from its router," *arXiv preprint arXiv:2410.12013*, 2024.
[61] B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Huang, J. Zhang, Y. Pang, M. Ning *et al.*, "MoE-LLaVA: Mixture of experts for large vision-language models," *arXiv preprint arXiv:2401.15947*, 2024.
[62] X. Fang, Y. Yang, and Y. Fu, "Visible-infrared person re-identification via semantic alignment and affinity inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 270–11 279.
[63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
[64] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[65] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

Xueting Chen received the M.S. degree in Artificial Intelligence from Xiamen University, China, in 2025. She is currently working toward the Ph.D. degree at the National University of Defense Technology, China. Her research interests include computer vision and pattern recognition.



Yan Yan (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, China, in 2009. He worked as a Research Engineer with the Nokia Japan Research and Development Center from 2009 to 2010. He worked as a Project Leader with the Panasonic Singapore Laboratory in 2011. He is currently a Full Professor with the School of Informatics, Xiamen University, China. He has published around 100 papers in the international journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, CVPR, ICCV, ECCV, AAAI, and ACM MM. His research interests include computer vision and pattern recognition.

Jing-Hao Xue (Senior Member, IEEE) received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is currently a Professor with the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He received the Best Associate Editor Award of 2021 from the IEEE Transactions on Circuits and Systems for Video Technology, and the Outstanding Associate Editor Award of 2022 from the IEEE Transactions on Neural Networks and Learning Systems.



Chang Shu received the Ph.D. degree in information and communication engineering from Tsinghua University, China, in 2011. He is currently a Lecturer with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, China. His research interests include computer vision and pattern recognition.

Hanzi Wang (Senior Member, IEEE) is currently a Distinguished Professor of "Minjiang Scholars" in Fujian province and a Founding Director of the Center for Pattern Analysis and Machine Intelligence at Xiamen University, China. He received his Ph.D. degree in Computer Vision from Monash University, where he was awarded the Douglas Lampard Electrical Engineering Research Prize and Medal for the best Ph.D. thesis. His research interests include computer vision and pattern recognition.