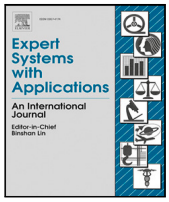
Expert Systems With Applications 251 (2024) 123885

Contents lists available at ScienceDirect

Expert Systems With Applications

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



HMRM: Hierarchy-aware Misclassification Risk Minimization for few-shot learning

Jie Jin, Yangqing Zhong, Hong Zhao *

School of Computer Science, Minnan Normal University, Zhangzhou, 363000, Fujian, China
Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou, Fujian, 363000, China

ARTICLE INFO

Keywords:
Hierarchical few-shot learning
Misclassification risk
Contrastive loss
Cost-sensitive learning

ABSTRACT

Hierarchical Few-shot Learning (HFSL) is a practical research of recognizing new categories from insufficient samples, which leverages multi-grained knowledge among samples to solve the data limitation problem. Current HFSL methods have delivered exceptional improvement and obtained outstanding performance. However, there are two potential and inevitable misclassification risk problems: Indifferent Risk occurs since most HFSL only concerns the correct classification of query samples and ignores the misclassifications; Indiscriminate Risk emerges in case of not distinguishing the error degree of these misclassifications. In this paper, we utilize multi-grained knowledge to propose Hierarchy-aware Misclassification Risk Minimization (HMRM) algorithm to minimize aforementioned two risks. HMRM consists of Cross Fine-grained Indifferent Risk Minimization (CFIRM) and Cross Coarse-grained Indiscriminate Risk Minimization (CCIRM) submodules. First, we propose CFIRM to design a fine-grained contrastive loss, which considers and evaluates misclassification results with fine-grained knowledge. Second, we present CCIRM to assign coarse-grained cost for misclassification via coarse-grained knowledge and cost-sensitive learning. Experimental results on FC100, CUB-200-2011, CIFAR-FS, and miniImageNet indicate the effectiveness of HMRM method. For instance, the accuracy of HMRM is at least 1.10% and 1.30% better than other methods in 5-way 1-shot on FC100 and CIFAR-FS. Our code is available at https://github.com/fhqxa/HMRM.

1. Introduction

Few-shot Learning (FSL) is a promising approach that recognizes new categories using only a few labeled examples (Li, Fergus, & Perona, 2006). The characteristic of FSL is to mimic the human learning process of new concepts, which can be rapidly adapted to new tasks by training only few data (Li, Jin, & Huang, 2022). However, insufficient data and expensive manual annotation costs hinder FSL tasks from acquiring greater remarkable effects (Xu, Wang, Chi, Yang, & Du, 2023). How to identify unlabeled samples from unseen categories with a small number of annotated samples remains a hard challenge (He, Xu, Shi, & Zhao, 2024; Lin, Shao, Zhou, Cai, & Liu, 2023).

Hierarchical Few-shot Learning (HFSL) can learn richer representations from limited data to overcome the challenge (Hu et al., 2023). HFSL studies different granularity characteristics of classes to learn more abundant sample relationships (Fu et al., 2023). From different granular perspectives, HFSL can be categorized into three main algorithms: fine-grained, coarse-grained, and multi-grained FSL methods (Zhao & Zhao, 2024). Fine-grained FSL unites fine granular features with deep learning network to enhance the effectiveness of FSL task (Xu, Zhang, Wei, & Wang, 2022; Zhu, Liu, & Jiang, 2020). Fine-grained knowledge can help the method distinguish subtle visual differences of classes (Fan, Bai, Sun, & Li, 2019). In the wake of superior performance in fine-grained recognition, easily acquired coarse knowledge is also applied in coarse-grained FSL approaches (Cui, Liao, Hu, An, & Liu, 2022; Saha, Cheng, & Maji, 2022). Independent of the above two single granularity FSL methods, multi-grained FSL methods reinforce constraints between the coarse and fine labels (Chen et al., 2023; Fu et al., 2023; Li, Luo, Lu, Xiang, & Wang, 2019; Liu, Zhou, Long, Jiang and Zhang, 2020), which improves the discrimination power of the framework for FSL task.

HFSL methods are effective and have yielded excellent results on FSL tasks. However, these researches overlook two potential misclassification risks: (1) indifferent risk exists since current HFSL counts much on cross-entropy loss, which only concerns the correct category classification and neglects remaining misclassification results; (2) indiscriminate risk emerges on condition that the method cares about the wrong classification results, but it does not distinguish the error degree of misclassification such as mean-squared loss. To further demonstrate

* Corresponding author at: School of Computer Science, Minnan Normal University, Zhangzhou, 363000, Fujian, China.
E-mail addresses: g2022061016@stu.mnnu.edu.cn (J. Jin), g2021061022@stu.mnnu.edu.cn (Y. Zhong), zh1127@mnnu.edu.cn (H. Zhao).

https://doi.org/10.1016/j.eswa.2024.123885
Received 11 October 2023; Received in revised form 27 March 2024; Accepted 30 March 2024
Available online 4 April 2024

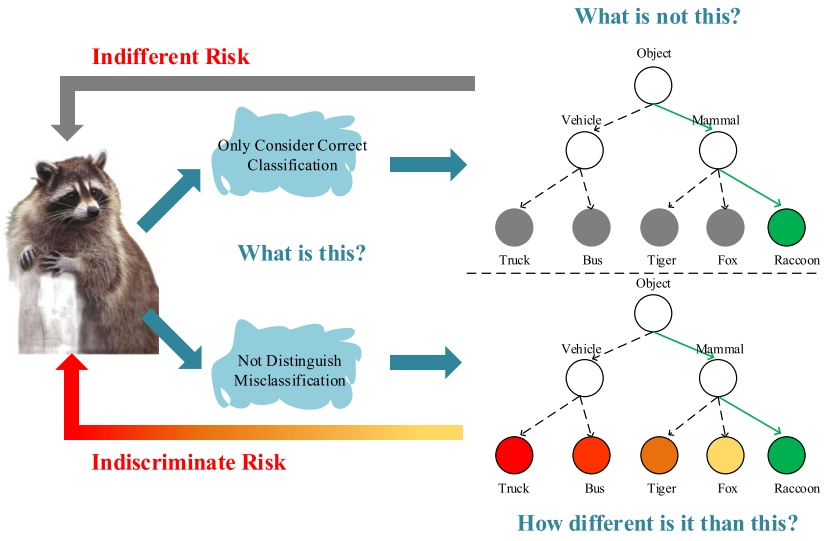


Fig. 1. A Raccoon example classification illustrates the indifferent risk and the indiscriminate risk. The green node is the ground truth, the gray node is neglected misclassification, and the remaining colors mean different error degrees of misclassification.

the existence of these two risks, we give an intuitive interpretation using a "Raccon" example classification in Fig. 1. Identifying a "Raccoon" sample as a different category is considered a misclassification. The FSL method will face indifferent risk if we only consider the prediction of the "Raccoon" and ignore the other four categories ("Truck", "Bus", "Tiger", and "Fox") predictions to measure this classification. The model needs to learn not only "What is this?" but also "What is not this?". The FSL method will face indiscriminate risk if we fail to distinguish the error degree of the four misclassifications. The model also needs to learn "How different is it than this?". For example, misclassifying "Raccoon" as "Truck" or "Bus" is a more serious error than misclassifying it as "Tiger" or "Fox". The misclassification results for a specific "Raccoon" sample are shown in Fig. 3. Misclassification case 1 has a probability of 0.2410 for the category "Raccoon" and 0.4612 for the category "Truck". Misclassification case 2 has a probability of 0.2521 for the category "Raccoon" and 0.7019 for the category "Fox". To solve the two risk problems, we propose Hierarchy-aware Misclassification Risk Minimization (HMRM) algorithm, which consists of Cross Fine-grained Indifferent Risk Minimization (CFIRM) and Cross Coarse-grained Indiscriminate Risk Minimization (CCIRM) submodules.

For the first indifferent risk problem, we propose the CFIRM submodule to pay more attention on misclassifications and impose penalties. We design a fine-grained contrastive loss to calculate the loss term of neglected misclassifications. Current contrastive loss reaches outperformance in feature embedding learning (Wang, Liu and Yu, 2021; Zhang, Peng, Zhang, & Wang, 2023), but we apply contrastive learning idea to probability distribution space of samples. Rather than dividing negative pairs by ground truth label, we integrate the query and support sample group in a misclassification as a negative pair. It can push the misclassification negative sample pairs further in probability distribution space. Besides, we utilize fine-grained relationships of negative pairs to replace the inner (dot) product calculation in conventional contrastive loss. The fine-grained information of samples exists extensively in FSL learning (Fan et al., 2019). It represents the semantic structure of samples and reduces feature matrix manipulation time consumption (Van Gansbeke, Vandenhende, Georgoulis, Proesmans, & Van Gool, 2020).

For the second indiscriminate risk problem, we present the CCIRM submodule to distinguish different error degrees of misclassifications. Motivated by the great performance of cost-sensitive learning (Cao, Zhao, & Zaiane, 2013; Liu, Qi, Xu, Gao, & Liu, 2019; Wang, Wang, Cheng, Li, & Zhang, 2019), we produce a coarse-grained cost of misclassification by learning coarse granular structural relationship and prediction probability of each negative pair. The coarse granular relationship is a guide of fine semantic structure, it can further measure negative pair distance in sample distribution space. The wrong prediction probability is another judgment for a misclassification. We give heavier cost for the misclassification as its coarser granular relationship and larger wrong prediction probability to minimize the indiscriminate risk.

We provide experiments on a set of datasets that demonstrate the remarkable performance and generalization ability of our HMRM algorithm. For example, HMRM has achieved at least 1.1% better than other models on FC100 and obtained 0.65%, 0.40%, and 0.37% improvement on CUB-200-2011, CIFAR-FS, and miniImageNet respectively at 5-way 1-shot setting. The main contributions of HMRM are listed as follows:

- We design a fine-grained contrastive loss to consider ignored misclassifications and impose penalties to minimize the indifferent risk. We select query and support sample groups in a misclassification as negative pairs and collect fine granular relationships of them to produce fine-grained contrastive loss terms.
- We present a coarse-grained cost of each misclassification and successfully distinguish the error degree of different misclassifications to minimize the indiscriminate risk. The coarser granular relationship and the larger the wrong prediction probability they are, the heavier cost we assign to misclassification result.
- Experiments on four popular FSL datasets certify that our HMRM algorithm outperforms the baseline method Relational Embedding Network (RENet) and reaches remarkable accuracy on the FSL task. We also investigate that our method is lightweight and both CFIRM and CCIRM can be utilized as plug-and-play submodules for kinds of FSL tasks.

The rest of this paper is settled as follows. Section 2 summarizes some related work briefly. Section 3 reports preliminary knowledge for this paper. Section 4 shows the implementation of HMRM algorithm and how to address the indifferent and indiscriminate risk problem. Experimental results on various datasets are presented in Section 5. Finally, we describe the conclusions and make arrangements for future work in Section 6.

2. Related work

We review related literatures on hierarchical few-shot learning (FSL), contrastive learning, and cost-sensitive learning.

2.1. Hierarchical few-shot learning

Hierarchical classification has been applied to FSL tasks. Hierarchical FSL can learn more abundant representations from limited labeled data by leveraging the similarities and shared characteristics between classes at different levels of granularity (Fu et al., 2023). From different granularity perspectives, Hierarchical FSL can be categorized into three primary algorithms, namely fine-grained, coarse-grained, and multi-grained methods (Zhao & Zhao, 2024).

Fine-grained FSL can recognize various subordinate classes by subtle and local differences (Zhu et al., 2020). Xu et al. (2022) present a dual attention architecture to extract fine-grained tailored feature representations for FSL. Similarly, Fan et al. (2019) utilize a large-margin prototype network with fine-grained features to enhance the generalization capability of FSL models. Inspired by the superior performance of fine-grained FSL, easier-found coarse-grained FSL approaches also contain abundant semantic knowledge (Cui et al., 2022). Saha et al. (2022) exploit coarse-grained labels to improve few-shot part segmentation models, which are easier and more available to obtain than per-pixel part labels for some categories. Unlike the above single-grained FSL methods, multi-grained FSL learns the relationship between coarse and fine-grained structure to reassimilate their acquired knowledge of probability distribution at two granularity levels to ensure accurate classification (Fu et al., 2023). Specifically, Li et al. (2019) leverage hierarchical information to train a multi-grained classifier to obtain an accurate classification for FSL tasks. Similarly, Liu, Zhou et al. (2020) collects coarse-grained knowledge to lessen the classification range of the fine-grained classes.

Although hierarchical structural knowledge greatly helps FSL tasks, many of these hierarchical FSL researches overlook indifferent and indiscriminate risk. Referring to the outstanding effects of hierarchical multi-grained knowledge in FSL tasks, we leverage strong constraints between coarse and fine granularity to minimize the aforementioned risks.

2.2. Contrastive learning

Contrastive learning has emerged as a fundamental element in self-supervised learning methodologies employed in image recognition (Kalantidis, Sariyildiz, Pion, Weinzaepfel, & Larlus, 2020). There are still two obstacles to implementing contrastive learning in FSL. First, the positive and negative sample selection strategy of traditional contrastive learning is at sample perspective (Saunshi, Plevrakis, Arora, Khodak, & Khandeparkar, 2019). They select augmented samples from a single sample as positive pairs, and all others are negative. Most existing contrastive learning methods might have a class collision issue regarding different samples from the same class as negative pairs (Hu, Wang, Hu, & Qi, 2021). Second, an extra pre-training process is regularly necessary for well-separated embedding space (Van Gansbeke et al., 2020), which adds additional computing burden and operational costs. Class collision and extra pre-training are obstructive factors for FSL tasks.

Therefore, we design a fine-grained contrastive loss to calculate the loss term of misclassifications (Wang, Liu and Yu, 2021; Zhang et al., 2023), which overcomes the two obstacles and minimizes the indifferent risk. We integrate the query and support sample group as positive pairs as long as they belong to the same fine granularity to settle the class collision issue. Uniting the fine-grained contrastive loss computation stage with FSL episodic training also avoids extra pre-training time consumption.

2.3. Cost-sensitive learning

Cost-sensitive learning is a resultful algorithm to address different risk problems in FSL classification, which can be divided into directly assigning costs and adaptive costs (Elkan, 2001). Directly assigning different costs for misclassifications is straightforward and effective (Cao et al., 2013; Zhao & Yu, 2019). For instance, Wang et al. (2019) introduce fuzzy memberships, which contain the different costs for each class, to consider different contributions to the decision boundary of different samples. Liu et al. (2019) involve algorithm modification by incorporating different costs for classification errors, which is particularly well-received in the context of Support Vector Machines. Instead of fixed cost for classification, adaptive cost-sensitive learning dynamically adjusts the cost weights for different class classifications. Xing, Lei, Yang, and Lu (2021) adaptively give greater weighting costs to shared filter kernels, which realize adaptive knowledge transfer among FSL tasks under the constraint of the changeable weighting costs. Analogously, Ren et al. (2022) propose an approach that considers the sample distribution, class convergence trends, and sample convergence trends to compute adaptive sample costs dynamically.

Inspired by the above two methods, we produce coarse-grained cost to minimize indiscriminate risk via integrating the two approaches mentioned above. We assign specified punishment weight directly for coarser-grained misclassification. Additionally, we consider the magnitude of the classification results to provide adaptive classification cost.

3. Preliminaries

We describe preliminary work on problem formulation, supervised contrastive learning, and relational embedding network methods.

3.1. Problem formulation

Few-shot Learning (FSL) is utilized to settle the issue of insufficient classification accuracy when there is limited data available for the FSL task. The former quantities demonstrated that Meta-learning has become a practical approach toward few-shot image recognition. Compared with the traditional deep learning method of "learning from scratch", meta-learning employs episode training as a training methodology. Each episode typically consists of a meta-task, which includes a predetermined number of samples from a set of categories. The Meta task adopts the N-way K-shot approach, which means N categories and K support samples of each category. It obtains required support set $S = \left\{(x_j^{(1)})_{j=1}^{K}, \ldots, (x_j^{(n)})_{j=1}^{K}, \ldots, (x_j^{(N)})_{j=1}^{K}\right\}$ and query set $Q = \left\{(x_i^{(1)})_{i=1}^{M}, \ldots, (x_i^{(n')})_{i=1}^{M}, \ldots, (x_i^{(N)})_{i=1}^{M}\right\}$ for each Meta task training. Here, M denotes the quantity of query samples in each category. The vector $x_j^{(n)}$ represents the jth support sample of the nth class, and the $x_i^{(n')}$ represents the ith query sample of the n'th class. Let y_i and $\hat{y}_i$ denote the ground truth class and prediction class of $x_i^{(n')}$. We can set a specific risk loss function $\mathcal{L}(g(\theta, x), y)$ to measure the misclassification risk, where $g(\cdot, \cdot)$ denotes the classifier for input (x, y) and θ represents the parameter of the classifier. The few-shot algorithm concentrates on finding parameter set θ^* to minimize the misclassification risk as follow:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(g(\theta, x), y). \tag{1}$$

3.2. Supervised contrastive learning

Supervised contrastive learning is a learning approach that intends to enhance the discerning ability of the model by contrasting positive and negative samples within the same class. The model learns to pull

similar samples closer in the feature space while pushing dissimilar samples apart. The loss takes the following form:

$$L_{out}^{sup} = \sum_{d \in D} L_{out,d}^{sup} = \sum_{d \in D} \frac{-1}{|B(d)|} \sum_{b \in B(d)} \log \frac{exp(z_d \cdot z_b/\tau)}{\sum_{c \in A(d)} exp(z_d \cdot z_c/\tau)}, \tag{2}$$

where $D = \{1, \ldots, 2N\}$ denotes the index of any augmented sample from N randomly sampled sample, and d, b, c are all parameters representing indexes of samples. The vector z_d is the representation feature of sample $x_d^{(n')}$. The $A(d) = D \setminus \{d\}$ represents the index of the rest of the samples. $B(d) = \{b \in A(d) : y_d = \hat{y}_d\}$ represents the set of positive samples that belong to the same class as $x_d^{(n')}$. $|B(d)|$ represents the number of samples in the set $B(d)$.

3.3. Relational embedding network

Relational Embedding Network (RENet) (Kang, Kwon, Min, & Cho, 2021) learn the relational corresponding embedding of query and support samples. RENet extracts significant relational corresponding embedding from "what to observe" and "where to attention" perspective. For the input one query sample $x_i^{(n')}$ of n'th class and support samples $(x_j^{(n)})_{j=1}^K$ of nth class, RENet utilizes self-correlation and co-attention computation function, $f(\theta, \cdot, \cdot)$, to obtain the final relational embedding vector $q = \left\{f(\theta, x_i^{(n')}, x_j^{(n)})\right\}_{j=1}^K$ and $s = \left\{f(\theta, x_j^{(n)}, x_i^{(n')})\right\}_{j=1}^K$. There are K support samples for the nth class in the N-way K-shot episode training. The average of K query and support relational embedding vectors represent prototype embedding set $\bar{q}^{(n')} = \sum_{j=1}^K \frac{1}{K} f(\theta, x_i^{(n')}, x_j^{(n)})$ and $\bar{s}^{(n)} = \sum_{j=1}^K \frac{1}{K} f(\theta, x_j^{(n)}, x_i^{(n')})$. The predicted class of query sample is determined by cosine similarity $sim(\bar{q}^{(n)}, \bar{s}^{(n')})$ of two prototype embedding (Snell, Swersky, & Zemel, 2017). RENet utilizes cross-entropy loss function $\mathcal{L}_{\text{metric}}$ to force each query relational embedding close to the support prototype embedding of the same class to learn better parameter set θ^* of the convolutional neural network:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{metric}} = \arg\min_{\theta} -\log \frac{\exp\left(\text{sim}\left(\bar{q}^{(n)}, \bar{s}^{(n)}\right)/\tau\right)}{\sum_{n'=1}^N \exp\left(\text{sim}\left(\bar{q}^{(n')}, \bar{s}^{(n')}\right)/\tau\right)}. \tag{3}$$

4. Hierarchy-aware misclassification risk minimization model

In this section, we describe our model in detail and divide it into three parts as shown in Fig. 2: (1) Cross Fine-grained Indifferent Risk Minimization (CFIRM) submodule designs a fine-grained contrastive loss based on misclassification results; (2) Cross Coarse-grained Indiscriminate Risk Minimization (CCIRM) submodule produces distinctive coarse-grained cost for each misclassification; (3) Hierarchy-aware Misclassification Risk Minimization (HMRM) model unites the fine-grained contrastive loss and assigned coarse-grained costs together to simultaneously minimize the indifferent and indiscriminate risk. In addition, we also address model analysis and learning to evaluate our model.

4.1. CFIRM: Cross fine-grained indifferent risk minimization

The main purpose of this module is to minimize indifferent risk. We utilize fine-grained knowledge to design a fine-grained loss to make it.

Indifferent Risk. Indifferent risk exists since most current few-shot learning counts much on cross-entropy loss, only concerning the correct classification of query sample and neglecting the misclassification. Let $p_{i,j}^{(n)}$ be the probability of classifying query sample $x_i^{(n')}$ into the nth class of support samples $(x_j^{(n)})_{j=1}^K$. Let $y_{i,j}^{(n)}$ be the one-hot label between query sample $x_i^{(n')}$ and support samples $(x_j^{(n)})_{j=1}^K$, which is defined as follow:

$$y_{i,j}^{(n)} = \begin{cases} 1, & \text{if } n = n' \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Hence, the cross-entropy loss $\mathcal{L}_{\text{ce}}$ computation is:

$$\begin{aligned} \mathcal{L}_{\text{ce}} &= \frac{1}{N} \sum_{n=1}^N y_{i,j}^{(n)} \log p_{i,j}^{(n)} \\ &= \frac{1}{N} \left\{ y_{i,j}^{(1)} \log p_{i,j}^{(1)} + \cdots + y_{i,j}^{(n')} \log p_{i,j}^{(n')} + \cdots + y_{i,j}^{(N)} \log p_{i,j}^{(N)} \right\} \\ &= \frac{1}{N} \log p_{i,j}^{(n')}. \end{aligned} \tag{5}$$

There is an indifferent risk that the loss function evaluates the model by considering one correct classification result rather than all categorical classification results. We leverage an example shown in Fig. 3 to intuitively explain the indifferent risk of cross-entropy loss. Case 1 misclassifies the "Raccoon" sample as "Truck", and case 2 misclassifies the "Raccoon" sample as "Fox". "Truck" is a kind of vehicle. "Raccoon" and "Fox" are both mammals. In accordance with the human learning process of new concepts, misclassification case 2 is more acceptable than misclassification case 1. Therefore, it is assumed that the loss term in Case 1 is suggested to be greater than the loss term in Case 2, and these two loss terms are the supposed loss terms. On the contrary, the

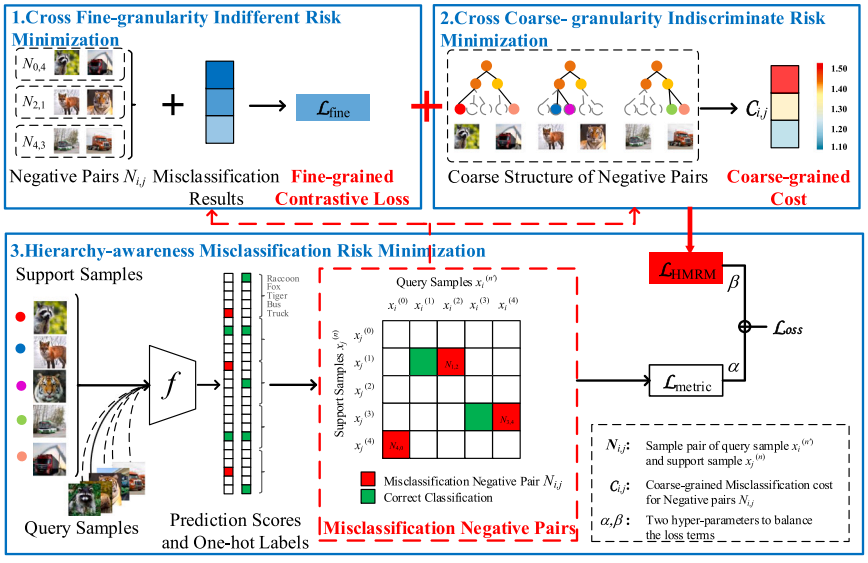


Fig. 2. The overall architecture of our proposed HMRM model. The symbol f denotes the few-shot learning mode.

cross-entropy loss term of case 1 is equal to that of case 2. Indifferent Risk exists due to this contradiction between reality and common sense.

Misclassification Contrastive Pairs Selection. The main approach of CFIRM is to design a fine-grained contrastive loss to minimize indifferent risk. According to Section 3.2, selecting negative and positive pairs is the precondition of contrastive loss. Instead of dividing sample pairs by ground truth label, we creatively select contrastive pairs based on classification results of each N-way K-shot episode training task. The negative pairs are drawn from misclassification query samples and corresponding support samples, and the rest correct classification pairs are positive pairs as follow:

$$NEG = \sum_{n=1}^{N} E(n) = \left\{ e \in E(n) : \hat{y}_e = n, y_e \neq n \right\}, \tag{6}$$

$$POS = \sum_{n=1}^{N} R(n) = \left\{ r \in R(n) : \hat{y}_r = n, y_r = n \right\}, \tag{7}$$

where NEG and POS contain index e and r of negative query sample $x_e^{(n')}$ and positive query sample $x_r^{(n')}$, and $E(n)$ denotes the sample set that contains the sample if its predicted label is n but its ground truth label is not n; similarly, $R(n)$ denotes another sample set that contains the sample if its predicted label and ground truth label both are n.

Fine-grained Contrastive Loss for Misclassification. In the N-way K-shot episode training task, there are $M \times N$ query sample classification results. Traditional work mostly relies on the cross-entry function to calculate this training task loss term by only adopting correct classification. The remaining misclassification results are not considered in loss computation, which causes the indifferent risk problem. CFIRM designs a fine-grained contrastive loss $\mathcal{L}_{\text{fine}}$ to assist existing loss to overcome this risk problem.

We utilize similarity $sim(\bar{q}^{(n')}, \bar{s}^{(n)})$ of negative pair to represent the fine granular relationship of this misclassification. The proposed loss function $\mathcal{L}_{\text{fine}}$ leverages the contrastive learning theory to pull positive correct classification pairs closer in the feature space while pushing negative misclassification pairs away. We can use the formula based on Eq. (2) to define the fine-grained contrastive loss:

$$\mathcal{L}_{\text{error}} = \sum_{n=1}^{N} \frac{1}{|E(n)|} \sum_{e \in E(n)} \log exp(sim(\bar{q}^{(n_e)}, \bar{s}^{(n)})/\tau), \tag{8}$$

$$\mathcal{L}_{\text{right}} = \sum_{n=1}^{N} \frac{-1}{|R(n)|} \sum_{r \in R(n)} \log exp(sim(\bar{q}^{(n_r)}, \bar{s}^{(n)})/\tau), \tag{9}$$

$$\mathcal{L}_{\text{all}} = \sum_{n=1}^{N} \frac{-1}{|R(n)|} \sum_{r \in R(n)} \log \frac{exp(sim(\bar{q}^{(n_r)}, \bar{s}^{(n)})/\tau)}{\sum_{i \in E(n) \cup R(n)} exp(sim(\bar{q}^{(n_i)}, \bar{s}^{(n)})/\tau)}, \tag{10}$$

where n_e is the class label of negative query sample $x_e^{(n_e)}$, n_r is the class label of positive query sample $x_r^{(n_e)}$, and n_i is the class label of query sample $x_i^{(n_i)}$. The three different loss formulations are not equivalent. The $\mathcal{L}_{\text{error}}$ pays attention on ignored misclassification results. However, $\mathcal{L}_{\text{right}}$ and $\mathcal{L}_{\text{all}}$ also consider positive correct classification results, which have already been put into computation by cross-entropy loss. The $\mathcal{L}_{\text{error}}$ is the most suitable auxiliary loss among the three losses. The final fine-grained contrastive loss is:

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{error}} = \sum_{n=1}^{N} \frac{1}{|E(n)|} \sum_{e \in E(n)} \log exp(sim(\bar{q}^{(n_e)}, \bar{s}^{(n)})/\tau). \tag{11}$$

4.2. CCIRM: Cross coarse-grained indiscriminate risk minimization

The main purpose of this module is to minimize indifferent risk. We utilize coarse granularity knowledge to create a coarse-grained cost to measure the error degree of misclassifications.

Indiscriminate Risk. After resolving any indifferent risks, addressing the immediately following indiscriminate risk becomes an urgent matter. Indiscriminate risk emerges when the model cannot distinguish the error degree of different misclassifications. The underlying assumption of mean squared loss is that the risks are the same for all wrong classifications. The mean-squared loss $\mathcal{L}_{\text{mse}}$ computation is:

$$\begin{aligned}
\mathcal{L}_{\text{mse}} &= \frac{1}{N} \sum_{n=1}^{N} (p_{i,j}^{(n)} - y_{i,j}^{(n)})^2 \\
&= \frac{1}{N} \left\{ (p_{i,j}^{(1)} - y_{i,j}^{(1)})^2 + \cdots + (p_{i,j}^{(n')} - y_{i,j}^{(n')})^2 + \cdots + (p_{i,j}^{(N)} - y_{i,j}^{(N)})^2 \right\} \\
&= \frac{1}{N} \left\{ (p_{i,j}^{(1)})^2 + \cdots + (p_{i,j}^{(n'-1)})^2 + (p_{i,j}^{(n')} - 1)^2 + (p_{i,j}^{(n'+1)})^2 + \cdots \right. \\
&\quad \left. + (p_{i,j}^{(N)})^2 \right\}.
\end{aligned} \tag{12}$$

The loss term obtains its minimum value at $p_{i,j}^{(1)} = \cdots = p_{i,j}^{(n'-1)} = p_{i,j}^{(n'+1)} = \cdots = p_{i,j}^{(N)}$. This means that mean-squared loss potentially assumes that all categories are equal with minimal differentiation. However, this assumption is not accordant with the real-world and has an indiscriminate risk. There are notable disparities among the results of the varied classifications.

We give an example shown in Fig. 3 to give an intuitive explanation of the indiscriminate risk. Case 1 misclassifies the "Raccoon" sample as "Truck", and case 2 misclassifies the "Raccoon" sample as "Fox". The distance between "Raccoon" and "Truck" classes is farther than the distance between "Raccoon" and "Fox" classes in the hierarchy. The latter case 2 is more acceptable than the former case 1. The loss term of case 1 should be larger than the loss term of case 2. In contrast, the mean-squared loss term of case 1 is smaller than the supposed loss term of case 2. The indiscriminate risk arises due to the conflict that the mean-squared loss cannot accurately adjust distinct weights for misclassifications rather than equal ones.

Fine-to-Coarse Granularity Structure Construction. The multi-grained structure plays a crucial role in few-shot learning based on its

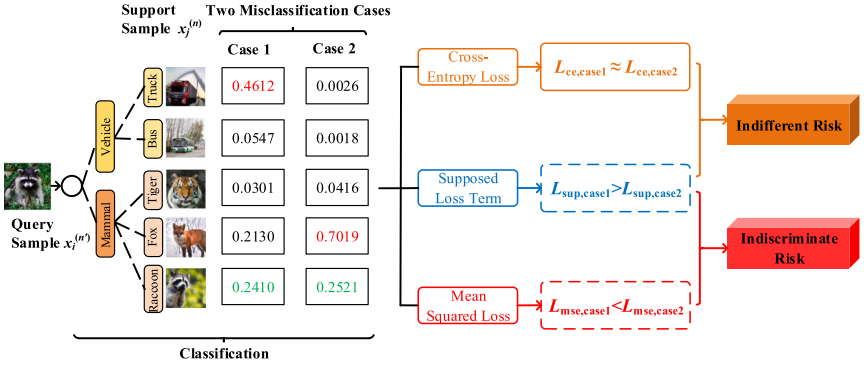


Fig. 3. Two different misclassification cases illustrate the indifferent and indiscriminate risks. Case 1 misclassifies the "Raccoon" sample as "Truck"; case 2 misclassifies the "Raccoon" sample as "Fox". The number in the rectangle is the predicted probability $p_{i,j}^{(n)}$; the red number is the maximum predicted probability, and the green number is the correct class predicted probability.

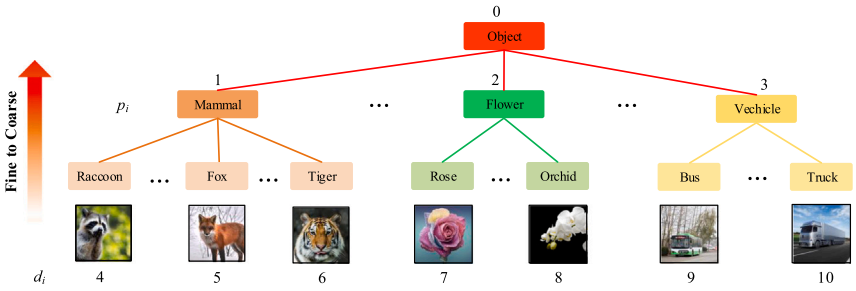


Fig. 4. The fine-to-coarse granularity structure of data example; d_i denotes fine nodes of samples; p_i denotes parent (coarse) node of d_j node.

semantic granular relationships. It allows for leveraging prior knowledge and efficiently utilizing limited labeled examples to classify novel classes. Higher-level or coarser granularity knowledge can provide semantic punishment costs for lower-level or finer granularity classification. The common training samples provided for meta-learning tasks merely have fine-grained labels, such as "Raccoon", "Fox", and "Bus". In the domain of meta-learning, the presence of similar features or patterns among different categories poses a challenge for fine-grained classification. The coarse granularity of these classes can offer additional guidance for FSL classification. Therefore, we adopt the semantic clustering method and WordNet (Miller, 1995; Wang, Liu, Lin et al., 2021) knowledge to construct the fine-to-coarse granularity structure as shown in Fig. 4.

Coarse-grained Sensitive Cost for Misclassification. As described before, the fine-to-coarse granularity structure provides coarse granular relation between query and support samples to evaluate misclassifications. It can also help the model measure the degree of misclassification. We introduce tree-induced error (*TIE*) (Zhao, Wang, Hu, & Zhu, 2019) to assess the hierarchical distance of query and support sample in a misclassification as follow:

$$TIE(d_i, d_j) = |E(d_i, d_j)|, \tag{13}$$

where d_i denotes the label of query sample $x_i^{(n')}$, and d_j denotes the label of support sample $x_j^{(n)}$. The $E(d_i, d_j)$ means path set from d_i to d_j, and $|\cdot|$ is the number of sets. We propose a function $\mathcal{H}(TIE(d_i, d_j), \gamma)$ to assess misclassification via concerning tree-induced error and computing a hierarchical distance punishment as follow:

$$\mathcal{H}(TIE(d_i, d_j), \gamma, \epsilon) = \begin{cases} 1, & \text{if } TIE(d_i, d_j) < \epsilon \\ \gamma, & \text{otherwise,} \end{cases} \tag{14}$$

where γ ($\gamma > 1.0$) represents a defined punishment hyper-parameter for misclassification if $TIE(d_i, d_j)$ is greater than hyper-parameter ϵ. The hierarchical distance punishment is positively correlated with the error degree of misclassification.

However, it is limited only concerning *TIE* to evaluate misclassification comprehensively. The prediction probability $p_{i,j}^{(n)}$ is also an important element to measure misclassification. Hence, we also propose another function $\mathcal{G}(p_{i,j}^{(n)}, \delta, \sigma)$ referring to adaptive cost (Li, Zhu, & Wang, 2023) to further assess misclassification punishment by generating a probability-sensitive cost as follow:

$$\mathcal{G}(p_{i,j}^{(n)}, \delta, \sigma) = \begin{cases} 1, & \text{if } p_{i,j}^{(n)} < \sigma \\ e^{-\delta \log\left(\frac{0.5+\sigma-p_{i,j}^{(n)}}{0.5-\sigma+p_{i,j}^{(n)}}\right)}, & \text{otherwise.} \end{cases} \tag{15}$$

The parameter δ is a smooth hyper-parameter to meet the prediction probability distribution of various categories. The parameter σ is a predictive probability evaluation criterion, we deploy higher cost in this misclassification as its probability $p_{i,j}^{(n)}$ is over σ.

With the support of the above two functions, the coarse-grained cost $C_{i,j}$ can be calculated by combining hierarchical distance punishment and probability-sensitive cost as follow:

$$C_{\text{coarse}} = \mathcal{H}(TIE(d_i, d_j), \gamma, \epsilon) \cdot \mathcal{G}(p_{i,j}^{(n)}, \delta, \sigma), \tag{16}$$

$$C_{\text{coarse}} = \begin{cases} 1, & \text{if } TIE(d_i, d_j) < \epsilon,\ p_{i,j}^{(n)} < \sigma \\ \gamma, & \text{if } TIE(d_i, d_j) \geq \epsilon,\ p_{i,j}^{(n)} < \sigma \\ e^{-\delta \log\left(\frac{0.5+\sigma-p_{i,j}^{(n)}}{0.5-\sigma+p_{i,j}^{(n)}}\right)}, & \text{if } TIE(d_i, d_j) < \epsilon,\ p_{i,j}^{(n)} \geq \sigma \\ \gamma e^{-\delta \log\left(\frac{0.5+\sigma-p_{i,j}^{(n)}}{0.5-\sigma+p_{i,j}^{(n)}}\right)}, & \text{otherwise.} \end{cases} \tag{17}$$

4.3. HMRM: Hierarchy-aware misclassification risk minimization

In this section, we discuss the theoretical foundation and a brief overview of HMRM methods.

Bayesian decision theory. Bayesian decision theory is a theoretical framework based on probability and decision theory (Wang et al., 2017), which is used for making optimal decisions in FSL. According to Bayesian decision theory, decision-makers introduce a loss function $\mathcal{L}_i$ to quantify the prediction cost and calculate the conditional risk $R(g(\theta, x_i^{(n')}), x_i^{(n')})$ for a meta-learning classification based on posterior probabilities $P_{i,n}$ of classifier $g(\theta, x_i^{(n')})$. They can be defined as follows:

$$\mathcal{L}_i = \begin{cases} 0, & \text{if } y_i = \hat{y}_i \\ 1, & \text{otherwise,} \end{cases} \tag{18}$$

$$R(g(\theta, x_i^{(n')}), x_i^{(n')}) = \sum_{n=1}^{N} \mathcal{L}_i P_{i,n}. \tag{19}$$

Hence, the best classifier is the one that achieves the minimum risk. The best parameter set θ^* of the model can be obtained by minimizing the conditional risk $R(g(\theta, x_i^{(n')}), x_i^{(n')})$ as:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} R(g(\theta, x_i^{(n')}), x_i^{(n')}) \\ &= \arg\min_{\theta} \sum_{n=1}^{N} \mathcal{L}_i P_{i,n}. \end{aligned} \tag{20}$$

However, there are still limitations in the Bayesian decision theory. It often assumes that all misclassification risks are equal as Eq. (18), which does not align with the actual situations. Indifferent and indiscriminate risks are all caused by this underlying rule. Therefore, HMRM presents a loss function $\mathcal{L}_{\text{HMRM}}$ to logically measure each misclassification risk by concerning the hierarchical relationship of query and support sample. The best parameter set θ^* is easily obtained by minimizing the misclassification loss function $\mathcal{L}_{\text{HMRM}}$ as:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{\text{HMRM}}. \tag{21}$$

According to Eq. (18), we understand that the evaluation of Empirical Risk Minimization (ERM) in traditional small sample learning is

Algorithm 1 HMRM: Hirearchy-aware Misclassification Risk Minimization

Input: The training set $D_{meta-train}$ in each episode, the number of classes N, the number of support examples K for each class; the number of query samples M for each class.

Output: The best parameter set θ^* of our model.

Iteration:

1: **for** each episode **do**
2: Randomly select samples to build support set S and query set Q;
3: Obtain relational embeddings and finish classification by Eq. (3);
4: Select *NEG* and *POS* pairs referring to classification results as Eq. (6) and Eq. (7);
5: Calculate fine-grained contrastive loss term $\mathcal{L}_{\text{fine}}$ by Eq. (11) based on *NEG* and *POS* sets;
6: Compute hierarchical distance punishment $\mathcal{H}(\cdot)$ and probability-sensitive cost $\mathcal{G}(\cdot)$ by Eq. (14) and Eq. (15);
7: Obtain the coarse-grained cost C_{coarse} via combining $\mathcal{H}(\cdot)$ and $\mathcal{G}(\cdot)$ by Eq. (17);
8: Unite $\mathcal{L}_{\text{fine}}$ with C_{coarse} to build a loss $\mathcal{L}_{\text{HMRM}}$ based on Eq. (23);
9: Update parameter θ of mode with $\mathcal{L}_{\text{HMRM}}$;
10: **end for**
11: **return** θ^*;

simple and poorly generalized, and the loss of misclassification results for prediction error is constant at 1. We design several loss functions to improve the robustness of the model and make the model more practical for ERM evaluation. Our new loss function not only scientifically evaluates the indifferent risk, but also provides an adaptive cost to discriminate risk to reduce the ERM.

Hierarchy-aware Misclassification Risk Minimization. HMRM is formed by combining fine-grained contrastive loss $\mathcal{L}_{\text{fine}}$ and coarse-grained cost C_{coare} as shown in Fig. 2. The final loss $\mathcal{L}_{\text{HMRM}}$ is a piecewise function, which is defined as follow:

$$\mathcal{L}_{\text{HMRM}} = C_{\text{coare}}\mathcal{L}_{\text{fine}}. \tag{22}$$

The best parameter set θ^* can be obtained by minimizing the final loss function by combining normal meta-learning loss $\mathcal{L}_{\text{metric}}$:

$$\begin{aligned}\theta^* &= \arg\min_{\theta} \left\{\alpha\mathcal{L}_{\text{HMRM}} + \beta\mathcal{L}_{\text{metric}}\right\} \\ &= \arg\min_{\theta} \left\{\alpha C_{\text{coare}}\mathcal{L}_{\text{fine}} + \beta\mathcal{L}_{\text{metric}}\right\},\end{aligned} \tag{23}$$

where parameters α and β are hyper-parameters which mean the weights of two loss terms.

4.4. Model analysis and learning

We propose the HMRM method by integrating multi-grained knowledge, contrastive learning, and cost-sensitive learning to minimize misclassification risks. The steps involved in the fine-tuning stage are outlined in Algorithm 1. Line 3 completes the classification via the baseline network, and then the negative and positive sample pairs are obtained in line 4. The fine-grained relationship is applied to assess ignored misclassification results in line 5, the coarse-grained structure information is utilized to distinguish different misclassifications in line 7. Line 8 combines fine-grained contrastive loss and coarse-grained cost of each misclassification to produce the final HMRM loss, which is utilized to update the parameter in line 9.

Compared with traditional FSL methods, HMRM has the following merits. First, it does not require the time-consuming pre-training stages and resource-heavy feature matrix operations as other supervised contrastive algorithms, which greatly decreases the computational complexity of the model. Second, it can utilize multi-grained knowledge of data to measure and distinguish different misclassification results. Third, it consists of two plug-and-play submodules, which can easily applied to many FSL scenarios. And our algorithm adopts a conventional few-shot deep learning framework, designing specific loss functions to reduce the risk of misclassification in model learning. The computational complexity of the model is at the same level as the commonly used cross entropy loss and mean squared loss functions. The widespread presence of two misclassification risks also demonstrates its significant generalization capability. To summarize, HMRM showcases a solid capability for low time consumption, plug-and-play functionality, and generalization. However, the limitation and drawback of the HMRM algorithm is that it requires high data structure. For tasks without hierarchical classification structure, the algorithm has limited performance improvement. More experiments are needed in the future to validate and improve it so that our model has better generalization ability.

5. Experiments

In this section, we use five metrics to evaluate the performance of the proposed HMRM algorithms: (1) performance comparison with the baseline and related methods; (2) performance analysis of three kinds of fine-grained contrastive losses; (3) ablation experiment analysis of CFIRM; (4) ablation experiment analysis of CCIRM; and (5) parameter sensitivity analysis.

5.1. Experiment settings

In our experiments, we concentrate on proving the effect of two submodules of HMRM and assessing it by plenty of comparison experiments with related FSL methods. We adopt ResNet-12 construction as the backbone network, self-correlation and co-attention are utilized for FSL. In accordance with the N-way K-shot evaluation protocol, we evaluate the model by conducting episodes with 15 query samples for each class. We report the average classification accuracy along with 95% confidence intervals derived from a random sample of 2000 test episodes. We deploy our model on four popular public FSL datasets: FC100 (Oreshkin, Rodríguez López, & Lacoste, 2018), CUB-200-2011 (Wah, Branson, Welinder, Perona, & Belongie, 2011), CIFAR-FS (Bertinetto, Henriques, Torr, & Vedaldi, 2018), and miniImageNet (Vinyals, Blundell, Lillicrap, Wierstra, et al., 2016). These datasets are particularly useful for assessing few-shot learning algorithms performance and adaptability to new categories.

5.2. Performance comparison with related methods

To investigate the effects of the HMRM method, the performance comparisons with related models on CUB-200-2011, CIFAR-FS, miniImageNet, and FC100 are implemented at the 5-way 1-shot and 5-shot settings. The encoder represents the architecture of the feature extractor and the best performance is highlighted. We can learn that the HMRM model has robust generalization ability on FSL tasks by comparing it with relevant few-shot methods on four datasets. HMRM has achieved a performance improvement of 0.65% improvement on CUB-200-2011 from Table 1, 1.3% improvement on CIFAR-FS from Table 2, 0.3% improvement on miniImageNet from Table 3, and 1.1% on the FC100 dataset from Table 4. This phenomenon demonstrates that our strategy of mitigating misclassification risks is advantageous for few-shot learning scenarios. The consistent improvement of experimental results showcases the robust generalization capability of our HMRM model, certifying its adaptability to diverse dataset configurations.

Similar to our HMRM, HyperShot (Sendera et al., 2023) suggests adjusting the classifier parameters according to the relationship between the sample embeddings. MTUNet2023 also utilizes a self-attention module to produce a distinguishable visual representation of samples

Table 1
Accuracy (%) comparison with different FSL methods at 5-way 1-shot and 5-shot experiment setting on CUB-200–2011.

Model	Encoder	5-way 1-shot	5-way 5-shot
RCN (Xue, Duan, Li, Chen, & Luo, 2020)	ResNet-12	74.65 ± 0.86	88.81 ± 0.57
FEAT (Ye, Hu, Zhan, & Sha, 2020)	ResNet-12	73.27 ± 0.22	85.77 ± 0.14
NegMargin (Liu, Cao et al., 2020)	ResNet-18	72.66 ± 0.85	89.40 ± 0.43
DeepEMD (Zhang, Cai, Lin, & Shen, 2020)	ResNet-12	75.65 ± 0.83	88.69 ± 0.50
MixtFSL (Afrasiyabi, Lalonde, & Gagné, 2021)	ResNet-10	73.94 ± 1.10	86.01 ± 0.50
HyperShot (Sendera et al., 2023)	Conv-4	66.13 ± 0.26	80.07 ± 0.22
RENet (Kang et al., 2021)	ResNet-12	79.21 ± 0.44	**91.10 ± 0.23**
RENet+CFIRM (ours)	ResNet-12	**79.86 ± 0.42**	90.90 ± 0.24

Table 2
Accuracy (%) comparison with different FSL methods at 5-way 1-shot and 5-shot experiment setting on CIFAR-FS.

Model	Encoder	5-way 1-shot	5-way 5-shot
ProtoNet (Snell et al., 2017)	ResNet12	53.31 ± 0.20	72.69 ± 0.74
RCN (Xue et al., 2020)	ResNet-12	69.02 ± 0.92	82.96 ± 0.67
MTUNet (Wang et al., 2023)	WRN-28-10	68.34 ± 0.49	82.93 ± 0.37
MTL (Wang, Zhao and Li, 2021)	ResNet-12	69.50 ± 0.30	84.10 ± 0.10
RENet (Kang et al., 2021)	ResNet-12	73.57 ± 0.46	86.75 ± 0.32
RENet+CFIRM (ours)	ResNet-12	74.13 ± 0.47	**86.87 ± 0.32**
RENet+CFIRM+CCIRM (ours)	ResNet-12	**74.88 ± 0.46**	86.59 ± 0.32

Table 3
Accuracy (%) comparison with different FSL methods at 5-way 1-shot and 5-shot experiment setting on miniImageNet.

Model	Encoder	5-way 1-shot	5-way 5-shot
ProtoNet (Snell et al., 2017)	ResNet12	62.39 ± 0.21	80.53 ± 0.14
MTUNet (Wang et al., 2023)	WRN-28-10	56.12 ± 0.43	71.93 ± 0.40
RCN (Xue et al., 2020)	ResNet-12	57.40 ± 0.86	75.19 ± 0.64
TADAM (Oreshkin et al., 2018)	ResNet12	58.50 ± 0.30	76.70 ± 0.30
FEAT (Ye et al., 2020)	ResNet-12	66.78 ± 0.20	82.05 ± 0.14
ConstellationNet (Xu, Wang, Tu, et al., 2020)	ResNet-12	64.89 ± 0.23	79.95 ± 0.17
NegMargin Liu, Cao et al. (2020)	ResNet-12	63.85 ± 0.81	81.57 ± 0.56
DeepEMD (Zhang et al., 2020)	ResNet12	65.91 ± 0.82	**82.41 ± 0.56**
MTL (Wang, Zhao et al., 2021)	ResNet-12	59.84 ± 0.22	77.72 ± 0.09
HyperShot (Sendera et al., 2023)	Conv-4	53.18 ± 0.45	69.62 ± 0.28
RENet (Kang et al., 2021)	ResNet-12	64.91 ± 0.44	80.29 ± 0.32
RENet+CFIRM (ours)	ResNet-12	**65.22 ± 0.44**	80.48 ± 0.32

Table 4
Accuracy (%) comparison with different FSL methods at 5-way 1-shot and 5-shot experiment setting on FC100.

Model	Encoder	5-way 1-shot	5-way 5-shot
TADAM (Oreshkin et al., 2018)	ResNet-12	40.10 ± 0.40	56.10 ± 0.40
MTL (Wang, Zhao et al., 2021)	ResNet-12	42.40 ± 0.20	57.70 ± 0.30
MABAS (Kim, Kim, & Kim, 2020)	ResNet-12	42.31 ± 0.75	57.56 ± 0.78
MixtFSL (Afrasiyabi et al., 2021)	ResNet-18	41.50 ± 0.67	58.39 ± 0.62
ConstellationNet (Xu et al., 2020)	ResNet-12	43.80 ± 0.20	59.70 ± 0.20
RENet (Kang et al., 2021)	ResNet-12	43.82 ± 0.41	60.19 ± 0.39
RENet+CFIRM (ours)	ResNet-12	44.47 ± 0.42	60.55 ± 0.39
RENet+CFIRM+CCIRM (ours)	ResNet-12	**44.92 ± 0.41**	**60.72 ± 0.40**

like our baseline model. Nevertheless, they ignore the risk of misclassification that hampers FSL applications in some tasks. CFIRM utilizes fine-grained knowledge of data to design a fine-grained contrastive loss to measure misclassification. FEAT (Ye et al., 2020) makes efforts to strengthen feature processing and choose the correct classification to produce loss terms. MixtFSL (Afrasiyabi et al., 2021) learns multi-modal mixed features for each class to extract a rich representation of samples. They attempt to acquire more enriched features to enhance the performance of FSL. However, it is worth noting that features derived from incorrectly classified samples are equally crucial for FSL. CCIRM can leverage the coarse-grained knowledge of misclassification samples to compute a coarse-grained cost to distinguish different misclassifications.

The results have shown that our method can compensate for the usual FSL approaches in terms of misclassification risk. The CFIRM and CCIRM consortium ensures that FSL has a stable option to mitigate the risk of misclassification. This modular composition of HMRM guarantees its flexibility, which also guarantees that we can choose different risk minimization strategies on several datasets with different granular structures.

5.3. Analysis of fine-grained contrastive loss

To assess the best fine-grained contrastive loss strategy of the CFIRM submodule, we conduct the comparison experiments on the CUB-200-2011 dataset at 5-way 1-shot setting. The contrastive-error loss $\mathcal{L}_{\text{error}}$ obtains better performance than the other two losses, as shown in Fig. 5. These three kinds of contrastive loss strategies all aim to minimize the indifferent risk of ignoring misclassification results. The $\mathcal{L}_{\text{error}}$ only focuses on the error classifications to make up for the deficiency of cross-entropy loss. It struggles to pay additional punishments for error classifications for the model. It can not only assist the model in memorizing relationships within the same category but also help strengthen the distance between different categories. On the contrary, the contrastive-right loss $\mathcal{L}_{\text{right}}$ loss aims to further strain ties of right classification relationships. It does not directly reflect the motivation

Table 5
Accuracy (%) comparison at 5-way 1-shot and 5-shot experiment settings on four datasets.

	1-shot		5-shot	
	RENet	RENet+CFIRM	RENet	RENet+CFIRM
FC100	43.82 ± 0.41	**44.47 ± 0.42**	60.19 ± 0.39	**60.55 ± 0.39**
miniImageNet	64.91 ± 0.44	**65.22 ± 0.44**	80.29 ± 0.32	**80.48 ± 0.32**
CIFAR-FS	73.57 ± 0.46	**74.13 ± 0.47**	86.75 ± 0.32	**86.87 ± 0.32**
CUB-200-2011	79.21 ± 0.44	**79.86 ± 0.42**	**91.10 ± 0.23**	90.90 ± 0.24

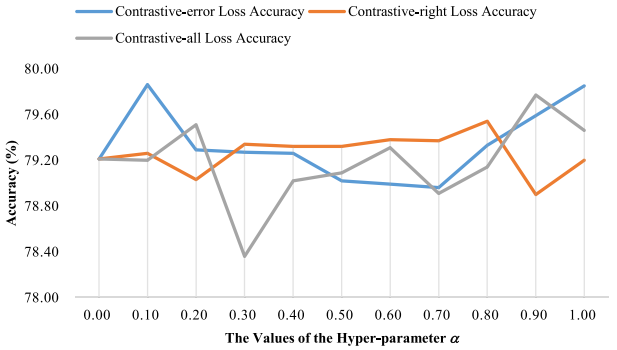


Fig. 5. Accuracy (%) comparison of three different fine-grained contrastive losses at 5-way 1-shot experiment setting on CUB-200-2011.

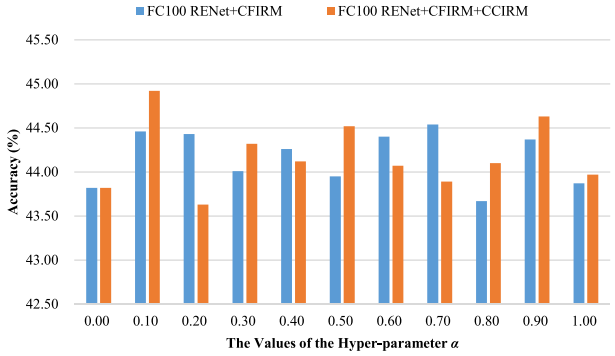


Fig. 6. Accuracy (%) comparison at 5-way 1-shot experiment setting on FC100.

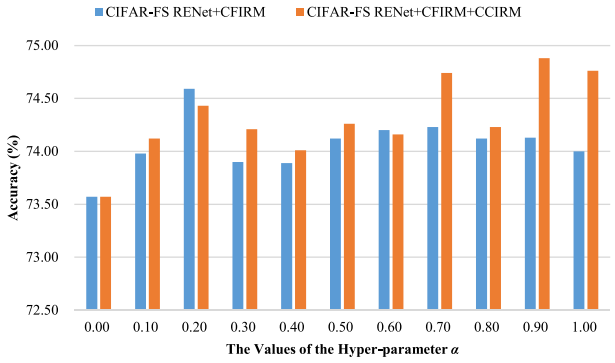


Fig. 7. Accuracy (%) comparison at 5-way 1-shot experiment setting on CIFAR-FS.

of the CFIRM model to focus on misclassification results. Besides, the contrastive-all loss $\mathcal{L}_{\text{all}}$ considers all results of correct and incorrect classification by treating correct classifications as positive sample pairs and incorrect classifications as negative sample pairs. Similar to supervised contrastive loss, it pushes negative pairs apart and pulls positive pairs closer together.

The $\mathcal{L}_{\text{all}}$ loss strategy is more reasonable, and its effectiveness surpasses the $\mathcal{L}_{\text{right}}$ loss strategy as depicted in Fig. 5. However, its repeated attention to correct classification results also conflicts with the existing loss functions in FSL. Based on this analysis, $\mathcal{L}_{\text{error}}$ is the strategy that aligns most with the expectations of the CFIRM model.

5.4. Ablation analysis of cross fine-grained indifferent risk minimization submodule

To evaluate the effectiveness of our CFIRM submodule in the FSL task, we deploy the comparison experiments on two methods: the baseline RENet with our CFIRM submodule and the baseline RENet. CFIRM leverages the fine-grained information inherent in the dataset and incorporates the concept of supervised contrastive learning to design fine-grained contrastive loss, which is introduced to the FSL task to reduce the associated misclassification risk and improve its performance. As described in Table 5, we can observe that the RENet with CFIRM method obtains better accuracy than the RENet method across four datasets at 5-way 1-shot and 5-shot settings. It is obvious that the CFIRM submodule helps RENet obtain 0.65% improvement on FC100, 0.30% improvement on miniImageNet, 0.56% improvement on CIFAR-FS, and 0.65% improvement on CUB-200-2011 at 5-way 1-shot setting. Similarly, the performance of CFIRM and RENet is also superior to that of RENet at the 5-way 5-shot setting on three datasets, with the exception of CUB-200-2011. CFIRM is effective primarily because it comprehensively considers all classification outcomes. Additionally, the effectiveness of CFIRM in enhancing fine-grained FSL capabilities is demonstrated by the high data similarity within these few-shot datasets.

5.5. Ablation analysis of cross coarse-grained indiscriminate risk minimization submodule

To further evaluate the effect of the CCIRM submodule, we first conduct the comparison experiments on two different methods: the baseline RENet with CFIRM submodule, the baseline RENet with CFIRM and CCIRM submodules. The experiments are at 5-way 1-shot setting on FC100 and CIFAR-FS as shown in Figs. 6 and 7. CFIRM selects negative and positive sample pairs and leverages the fine-grained relationship of samples to assess the specific misclassification. CCIRM collects the coarse-grained information of samples to define a coarse-grained cost for misclassification. The experimental results indicate that CFIRM and CCIRM submodules outperform the baseline at most settings. Ten different α values were used to test the effects of CFIRM and CCIRM. The results show that CFIRM outperforms baseline in the majority of experimental settings. The results show that CFIRM outperforms baseline in the majority of experimental settings. Additionally, CCIRM performs better than CFIRM in most experimental settings.

However, incorporating the CCIRM submodule can help the CFIRM submodule obtain greater performance at most experiment settings ($\alpha = \{0.10, 0.30, 0.50, 0.80, 0.90, 1.00\}$). Specifically, we add another baseline RENet method and compare experiment results on three different methods of 5-way 1-shot and 5-shot settings on FC100 and CIFAR-FS. As shown in Table 6, the accuracy of RENet and CFIRM method achieves 44.92% in 1-shot and 60.72% in 5-shot when combined with the CCIRM submodule on FC100. The CCIRM helps CFIRM reach 1.10% and 0.62% improvement in 1-shot and 5-shot, while the CFIRM obtains 0.65% and 0.44% improvement. Meanwhile, when the RENet and CFIRM methods were used in conjunction with the CCIRM submodule on CIFAR-FS, the accuracy reached 0.75% improvement in 1-shot setting.

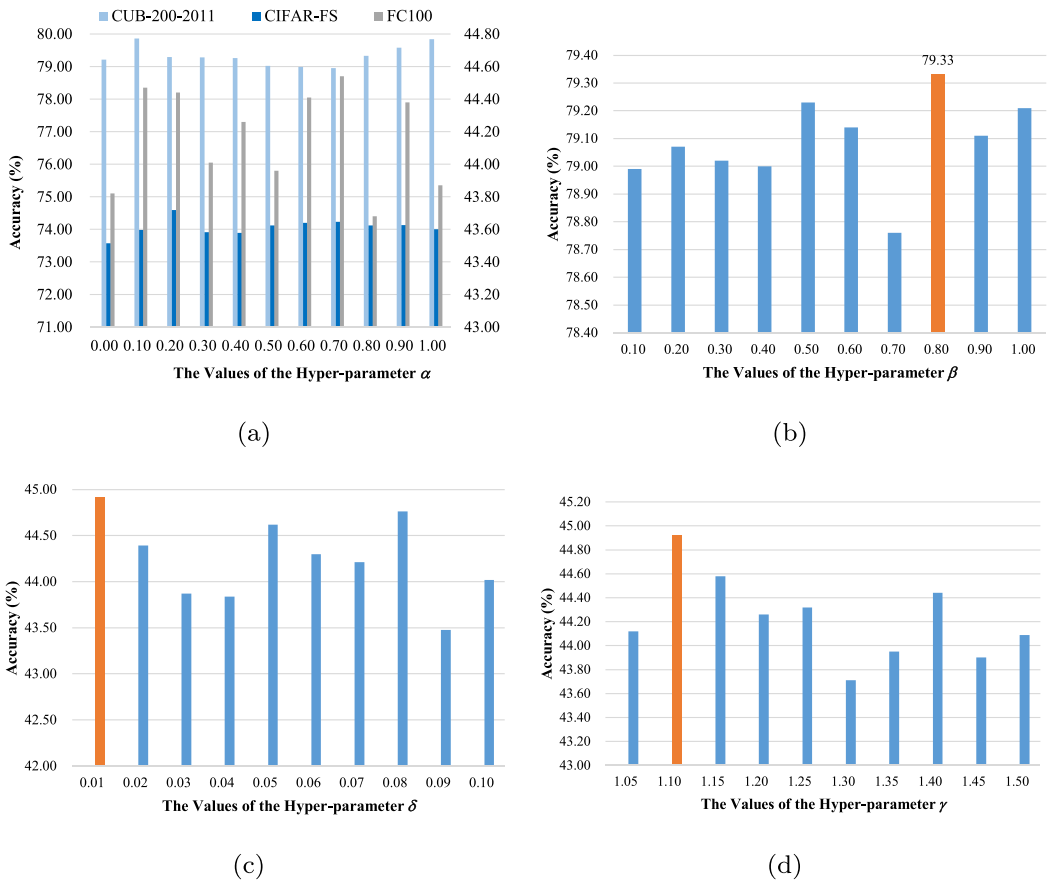


Fig. 8. The sensitivity experiment for hyper-parameters of our HMRM model at 5-way 1-shot setting.

Table 6
Accuracy (%) comparison at 5-way experiment setting on FC100 and CIFAR-FS.

	1-shot		
	RENet	RENet+CFIRM	RENet+CFIRM+CCIRM
FC100	43.82 ± 0.41	44.47 ± 0.42	**44.92 ± 0.41**
CIFAR-FS	73.57 ± 0.46	74.13 ± 0.47	**74.88 ± 0.46**
	5-shot		
	RENet	RENet+CFIRM	RENet+CFIRM+CCIRM
FC100	60.19 ± 0.39	60.55 ± 0.39	**60.72 ± 0.39**
CIFAR-FS	86.75 ± 0.32	**86.87 ± 0.32**	86.59 ± 0.32

The coarse-grained cost of CCIRM helps the CFIMR submodule distinguish various differences among misclassification categories and obtain better performance.

5.6. Parameters sensitivity analysis

We implement experiments on CUB-200-2011, CIFAR-FS, and FC100 to demonstrate the influence of varied hyper-parameters within our HMRM framework. Four hyper-parameters are analyzed at the 5-way 1-shot experimental setting.

- First, the hyper-parameter α denotes the weight of our fine-grained contrastive loss of CFIMR in the final FSL loss. The hyper-parameter α is analyzed on CUB-200-2011, CIFAR, and FC100, as shown in Fig. 8(a). The values of α are set to $\{0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00\}$. Note that our model achieves the best performance on CUB-200-2011 when setting $\alpha = 0.10$. Meanwhile, the experiment on CIFAR-FS obtain top performance at $\alpha = 0.10$ and $\alpha = 0.20$ settings. The experiment on FC100 reach top accuracy at $\alpha = 0.10$ and $\alpha = 0.70$ settings. After conducting a thorough analysis of the experimental results, we ultimately chose α value of 0.1.

- Second, the hyper-parameter β denotes the weight of original loss of FSL in the final loss. The hyper-parameter β experiment on the CUB-200-2011 dataset in Fig. 8(b) is deployed to discuss its effect, we set $(1 - \beta)$ to replace the value of α. Our model shows the best performance when setting $(\alpha = 0.20, \beta = 0.80)$. The accuracy on the CUB-200-2011 dataset reaches 79.33%. Both the above two settings of hyper-parameter α and β reach better performance than the baseline model without our CFIRM submodule, but the former $(\alpha = 0.10, \beta = 1.00)$ setting gets the best.

- Third, the hyper-parameter γ denotes the coarse-grained punishment for misclassification. The hyper-parameter γ is also explored on experiments in Fig. 8(c), it is set different values under the HMRM model to test the influence of model. Our HMRM model obtains the best performance when setting $\gamma = 1.10$ in Fig. 8(c). It is obvious that the performance of HMRM descends with coarse-grained punishment γ increasing.

- Finally, the hyper-parameter δ is positively correlated with the punishment cost of misclassification. We discuss the effectiveness of hyper-parameter δ. Similar to γ, our model achieves the best performance when setting $\delta = 0.01$ in Fig. 8(d), the performance of HMRM decreases while the coarse-grained penalty parameter γ ascends.

6. Conclusions and future work

In this paper, we propose a Hierarchy-aware Misclassification Risk Minimization (HMRM) model to resolve underlying misclassification

risks. To mitigate the indifferent risk that arises from ignoring misclassification outcomes, a fine-grained contrastive loss based on fine granular relationships is designed to define the loss terms for misclassification results. Furthermore, failure to distinguish the degree of errors among various misclassification outcomes poses an additional indiscriminate risk. We address this by defining a coarse-grained cost for each misclassification loss term based on coarse granular relationships and cost-sensitive learning. By combining the fine-grained contrastive loss and coarse-grained cost, we can not only focus on misclassification outcomes but also precisely define loss terms for different misclassification results. It assists the model in adapting more efficiently to the distribution of data and enhances its performance in classification. Experimental results on four datasets validate the effectiveness of the HMRM model, and the plug-and-play characteristic of HMRM further demonstrates its strong generalization capabilities. However, the effectiveness is limited for HMRM to design auxiliary loss to address the misclassification risk problem after the decision stage. The final result of classification is heavily influenced by the processing of data and features during the training and learning stage. In future work, we will concentrate on selecting optimal features through data processing methods to further decrease the risk of misclassification.

CRediT authorship contribution statement

Jie Jin: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Yangqing Zhong:** Validation, Data curation. **Hong Zhao:** Resources, Supervision, Investigation, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62376114 and the Natural Science Foundation of Fujian Province under Grant No. 2021J011003.

References

Afrasiyabi, A., Lalonde, J. F., & Gagné, C. (2021). Mixture-based feature space learning for few-shot image classification. In *IEEE/CVF International Conference on Computer Vision* (pp. 9041–9051).

Bertinetto, L., Henriques, J. F., Torr, P. H., & Vedaldi, A. (2018). Meta-learning with differentiable closed-form solvers. arXiv preprint arXiv:1805.08136.

Cao, P., Zhao, D., & Zaiane, O. (2013). An optimized cost-sensitive svm for imbalanced data learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 280–292). Springer.

Chen, Z., Fu, L., Yao, J., Guo, W., Plant, C., & Wang, S. (2023). Learnable graph convolutional network and feature fusion for multi-view learning. *Information Fusion*, *95*, 109–119.

Cui, Y., Liao, Q., Hu, D., An, W., & Liu, L. (2022). Coarse-to-fine pseudo supervision guided meta-task optimization for few-shot object classification. *Pattern Recognition*, *122*, Article 108296.

Elkan, C. (2001). The foundations of cost-sensitive learning. *Vol. 17*, In *International Joint Conference on Artificial intelligence* (pp. 973–978). Lawrence Erlbaum Associates Ltd.

Fan, M., Bai, Y., Sun, M., & Li, P. (2019). Large margin prototypical network for few-shot relation classification with fine-grained features. In *ACM International Conference on Information and Knowledge Management* (pp. 2353–2356).

Fu, Y., Wang, S., Li, X., Li, D., Li, Y., Liao, J., et al. (2023). Hierarchical neural network: Integrate divide-and-conquer and unified approach for argument unit recognition and classification. *Information Sciences*, *624*, 796–810.

He, Wenwei, Xu, Junyan, Shi, Jie, & Zhao, Hong (2024). ECS-SC: Long-tailed classification via data augmentation based on easily confused sample selection and combination. *Expert Systems with Applications*, [ISSN: 0957-4174] *246*, 123138.

Hu, G., He, W., Sun, C., Zhu, H., Li, K., & Jiang, L. (2023). Hierarchical belief rule-based model for imbalanced multi-classification. *Expert Systems with Applications*, *216*, Article 119451.

Hu, Q., Wang, X., Hu, W., & Qi, G. (2021). Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1074–1083).

Kalantidis, Y., Sariyildiz, M. B., Pion, N., Weinzaepfel, P., & Larlus, D. (2020). Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, *33*, 21798–21809.

Kang, D., Kwon, H., Min, J., & Cho, M. (2021). Relational embedding for few-shot classification. In *IEEE/CVF International Conference on Computer Vision* (pp. 8822–8833).

Kim, J., Kim, H., & Kim, G. (2020). Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning. In *European Conference Computer Vision 2020*, 599–617.

Li, F.-F., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*, 594–611.

Li, L., Jin, W., & Huang, Y. (2022). Few-shot contrastive learning for image classification and its application to insulator identification. *Applied Intelligence*, *52*, 6148–6163.

Li, A., Luo, T., Lu, Z., Xiang, T., & Wang, L. (2019). Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7212–7220).

Li, J., Zhu, X., & Wang, J. (2023). Adaboost.c2: Boosting classifiers chains for multi-label classification. *Vol. 37*, In *AAAI Conference on Artificial Intelligence* (pp. 8580–8587).

Lin, J., Shao, H., Zhou, X., Cai, B., & Liu, B. (2023). Generalized maml for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals. *Expert Systems with Applications*, Article 120696.

Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., et al. (2020). Negative margin matters: Understanding margin in few-shot classification. In *European Conference Computer Vision 2020*, 438–455.

Liu, N., Qi, E., Xu, M., Gao, B., & Liu, G. (2019). A novel intelligent classification model for breast cancer diagnosis. *Information Processing & Management*, *56*, 609–623.

Liu, L., Zhou, T., Long, G., Jiang, J., & Zhang, C. (2020). Many-class few-shot learning on multi-granularity class hierarchy. *IEEE Transactions on Knowledge and Data Engineering*, *34*, 2293–2305.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, *38*, 39–41.

Oreshkin, B., Rodríguez López, P., & Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in Neural Information Processing Systems*, *31*.

Ren, Z., Zhu, Y., Kang, W., Fu, H., Niu, Q., Gao, D., et al. (2022). Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data. *Knowledge-Based Systems*, *241*, Article 108296.

Saha, O., Cheng, Z., & Maji, S. (2022). Improving few-shot part segmentation using coarse supervision. In *European Conference on Computer Vision* (pp. 283–299). Springer.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., & Khandeparkar, H. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning* (pp. 5628–5637).

Sendera, M., Przewięźlikowski, M., Karanowski, K., Zięba, M., Tabor, J., & Spurek, P. (2023). Hypershot: Few-shot learning by kernel hypernetworks. In *IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2469–2478).

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., & Van Gool, L. (2020). Scan: Learning to classify images without labels. In *European Conference on Computer Vision* (pp. 268–285). Springer.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The caltech-ucsd birds-200-2011 dataset*. California Institute of Technology.

Wang, Y., Hu, Q., Zhou, Y., Zhao, H., Qian, Y., & Liang, J. (2017). Local bayes risk minimization based stopping strategy for hierarchical classification. In *2017 IEEE International Conference on Data Mining* (pp. 515–524). IEEE.

Wang, B., Li, L., Verma, M., Nakashima, Y., Kawasaki, R., & Nagahara, H. (2023). Match them up: Visually explainable few-shot image classification. *Applied Intelligence*, *53*, 10956–10977.

Wang, Y., Liu, R., Lin, D., Chen, D., Li, P., Hu, Q., et al. (2021). Coarse-to-fine: Progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, X., Liu, Z., & Yu, S. X. (2021). Unsupervised feature learning by cross-level instance-group discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12586–12595).

Wang, Z., Wang, B., Cheng, Y., Li, D., & Zhang, J. (2019). Cost-sensitive fuzzy multiple kernel learning for imbalanced problem. *Neurocomputing*, *366*, 178–193.

Wang, H., Zhao, H., & Li, B. (2021). Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *International Conference on Machine Learning* (pp. 10991–11002).

Xing, S., Lei, Y., Yang, B., & Lu, N. (2021). Adaptive knowledge transfer by continual weighted updating of filter kernels for few-shot fault diagnosis of machines. *IEEE Transactions on Industrial Electronics*, *69*, 1968–1976.

Xu, X., Wang, Z., Chi, Z., Yang, H., & Du, W. (2023). Complementary features based prototype self-updating for few-shot learning. *Expert Systems with Applications*, *214*, Article 119067.

Xu, W., Wang, H., Tu, Z., et al. (2020). Attentional constellation nets for few-shot learning. In *International Conference on Learning Representations*.

Xu, S., Zhang, F., Wei, X., & Wang, J. (2022). Dual attention networks for few-shot fine-grained recognition. *Vol. 36*, In *AAAI Conference on Artificial Intelligence* (pp. 2911–2919).

Xue, Z., Duan, L., Li, W., Chen, L., & Luo, J. (2020). Region comparison network for interpretable few-shot image classification. arXiv preprint arXiv:2009.03558.

Ye, H., Hu, H., Zhan, D., & Sha, F. (2020). Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8808–8817).

Zhang, C., Cai, Y., Lin, G., & Shen, C. (2020). Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12203–12213).

Zhang, X., Peng, H., Zhang, J., & Wang, Y. (2023). A cost-sensitive attention temporal convolutional network based on adaptive top-k differential evolution for imbalanced time-series classification. *Expert Systems with Applications*, *213*, Article 119073.

Zhao, H., Wang, P., Hu, Q., & Zhu, P. (2019). Fuzzy rough set based feature selection for large-scale hierarchical classification. *IEEE Transactions on Fuzzy Systems*, *27*, 1891–1903.

Zhao, H., & Yu, S. (2019). Cost-sensitive feature selection via the $\ell_{2,1}$-norm. *International Journal of Approximate Reasoning*, *104*, 25–37.

Zhao, W., & Zhao, H. (2024). Hierarchical long-tailed classification based on multi-granularity knowledge transfer driven by multi-scale feature fusion. *Pattern Recognition*, *145*, Article 109842.

Zhu, Y., Liu, C., & Jiang, S. (2020). Multi-attention meta learning for few-shot fine-grained image recognition. In *International Joint Conference on Artificial Intelligence* (pp. 1090–1096).