This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



# Few-shot Learning with Multi-Granularity Knowledge Fusion and Decision-Making

Yuling Su, Hong Zhao, *Member, IEEE,* Yifeng Zheng, and Yu Wang

***Abstract*—Few-shot learning (FSL) is a challenging task in classifying new classes from few labelled examples. Many existing models embed class structural knowledge as prior knowledge to enhance FSL against data scarcity. However, they fall short of connecting the class structural knowledge with the limited visual information which plays a decisive role in FSL model performance. In this paper, we propose a unified FSL framework with multi-granularity knowledge fusion and decision-making (MGKFD) to overcome the limitation. We aim to simultaneously explore the visual information and structural knowledge, working in a mutual way to enhance FSL. On the one hand, we strongly connect global and local visual information with multi-granularity class knowledge to explore intra-image and inter-class relationships, generating specific multi-granularity class representations with limited images. On the other hand, a weight fusion strategy is introduced to integrate multi-granularity knowledge and visual information to make the classification decision of FSL. It enables models to learn more effectively from limited labelled examples and allows generalization to new classes. Moreover, considering varying erroneous predictions, a hierarchical loss is established by structural knowledge to minimize the classification loss, where greater degree of misclassification is penalized more. Experimental results on three benchmark datasets show the advantages of MGKFD over several advanced models.**

***Index Terms*—Few-shot learning, multi-granularity, knowledge fusion, decision-making.**

## I. Introduction

FEW-shot learning (FSL) is a significant and hot topic in the field of machine learning, which aims to recognize new classes from few examples [1]. The availability of only one or very few examples challenges the generalization ability of machine learning. In 2000, Miller et al. [2] first proposed the problem of learning from very few examples. Thereafter, more and more efforts were devoted to the FSL research. Recently, FSL models are being well widely applied to various research fields such as computer vision, natural language processing, and data analysis [3], [4].

The key objective of FSL is to establish a connection between the knowledge learned from the base classes and apply it to effectively recognize new classes using low-shot data. Among them, metric learning is one of the mainstream approaches for FSL, which strives to find an optimal similarity measurement space to bridge the gap between base classes and new classes [5]. For instance, Matching Network [6], Prototypical Network [7], and Relation Network [8], classic FSL models, use the Cosine distance, Euclidean distance, or a learnable module to construct a common metric space of bases and new classes. In addition, Zhang et al. [9] found that directly computing the distance between two global features may be influenced by the cluttered background and large intra-class appearance variations. Thus, they leveraged the Earth Mover's distance to minimum matching cost between local features in two samples. Unlike modelling by a single-scale perspective, some scholars further study from a multi-scale perspective to excavate more potential information [10]–[12]. For example, Jiang et al. [10] extracted multi-scale features and learn the multi-scale relations between samples for FSL. Similarly, a Bi-similarity Network was proposed with two similarity measures to obtain discriminative feature maps [11]. These models intend to mine visual information within data whether from single- or multi-scales, but they cannot satisfy the development of FSL. They fall short of capturing and possessing the latent information in structural knowledge of data.

Structural knowledge of data commonly involves inherent relationships and dependencies between different classes, which can compensate for the data scarcity problem of FSL [13], [14]. The structural knowledge like hierarchical class structure demonstrates the multi-granularity correlations among the fine- and coarse-grained classes, serving as an external important knowledge to guide FSL [15]–[17]. For instance, Li et al. [17] leveraged the class hierarchy as prior knowledge to construct a coarse-to-fine FSL classifier. Similarly, Zhang et al. [18] adopted the hierarchical structure to design an interpretable decision tree-based classifier. However, though these models have achieved great efforts, they are failed to establish a coherent connection between visual data and structured knowledge, especially with available limited data. They focus on exploring how to embed structural information into FSL without fully taking advantage of the visual information of data, which plays equal roles to enhance FSL.

In this paper, we propose a new FSL model via multi-granularity knowledge fusion and decision-making (MGKFD), which connects the visual information and structural knowledge from image and class levels. We strive to simultaneously exploit the abundant visual information and structural knowledge, working in a mutual way for FSL. Specifically, MGKFD is mainly divided into two parts: multi-granularity feature extraction and multi-granularity knowledge fusion and decision-making. For an image, it explicitly represents class specific and detailed information, where the global feature is

Y. L. Su, H. Zhao, and Y. F. Zheng are with the School of Computer Science, Minnan Normal University, Zhangzhou, Fujian, 363000, China, and with the Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou, Fujian, 363000, China, (e-mail: Syuling@126.com, hongzhaocn@163.com,zyf@mnnu.edu.cn). (*Corresponding author: Hong Zhao.*)

Y. Wang is with the School of Intelligence and Computing, Tianjin University, Tianjin, 300072, China, e-mail: wang.yu@tju.edu.cn

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



a crude representation of image content and the local feature is better at capturing distinctive information on the objects in an image. Similarly, in a hierarchical class structure, the coarse-grained classes have higher generalization performance while the fine-grained classes generally have a lesser diversity of features. Thereby, we initially represent global features as coarse-grained class representations, while local features are considered as fine-grained representations. That allows us to investigate both intra-image and inter-class relationships, ultimately yielding distinct class representations with few-shot images. Then, we fuse the multi-granularity knowledge to guide the FSL classification decision-making. The representations and discriminative knowledge of coarse-grained classes are strongly connected to fine-grained classes, which makes them valuable for the learning of fine-grained classes. In this way, we introduce a fusion strategy to integrate the multi-granularity knowledge for final FSL decision-making. Additionally, we introduce a hierarchical loss to minimize classification error, which assigns different classification risks based on the misclassification degree guided by the structural knowledge. To sum up, we aim to explicitly employ the structured knowledge to integrate the limited visual information and ultimately improving the model performance for FSL tasks.

To validate the performance of MGKFD, we conduct the extensive experiments on several publicly data. The experimental results prove the effectiveness of our model against several advanced FSL models from different perspectives. The contributions of this paper are summarized as follows:

- We propose a new few-shot framework by explicitly connecting abundant class structural knowledge with limited visual information for FSL classification decision-making.
- We design a fusion strategy to integrate the multi-granularity knowledge, which can relieve the risk of inter-layer error transmission than a coarse-to-fine decision.
- A hierarchical loss is introduced to minimize the degree of classification error with different weights unlike general loss functions that assume that the misclassification errors are equal.

In the following sections, we present a brief review of related work in Section II. Then, the proposed model is introduced in Section III. Next, the experimental settings are introduced in Section IV. We report and analyse the experimental results in Section V. Finally, the conclusions of this paper and a further study are summarized in Section VI.

## II. RELATED WORK

In this section, we briefly introduce the work related to our study, including few-shot learning and few-shot learning based on structural knowledge.

### Few-shot learning

The FSL models are generally based on meta-learning and aim to learn transferable knowledge from only a few data to new tasks [21], and the existing models can be categorized into three types. First, data augmentation is an effective way to increase training samples or enhance data features [5]. It attempts to solve the problem of insufficient data by increasing the amount of data. Second, some scholars focus on the optimization of models for solving the FSL problem without data augmentation, called the optimization-based method. It aims to learn how to rapidly update models online given a small number of support samples, such as MAML [23] and MetaOptNet [23]. The third method is metric learning, and its objective is to learn the similarity between query and support images [8]. For instance, the Cosine and Euclidean distances were first adopted in FSL to measure image similarity [6], [7]. Also, Kang et al. [24] leveraged the relational patterns within and between images to measure their similarity. In addition, to avoid the influence of the cluttered background and large intra-class appearance variations, Zhang et al. [9] used Earth Mover's distance to measure the image similarity by the minimum matching cost. Unlike single-scale metric learning, many scholars further build FSL models from a multi-scale perspective [10]–[12]. For features, Jiang et al. [10] proposed a multi-scale relation generation network to learn the multi-scale relations between samples for FSL. From the measurement perspective, a Bi-similarity Network was proposed with two similarity measures to obtain discriminative feature maps [11]. From the perspectives of single- or multi-scales, these models focus on taking advantage of latent visual information within data, but they cannot exploit the latent structural knowledge among classes.

### Few-shot learning based on structural knowledge

The structural knowledge inherent in classes provides strong class semantic information, and many studies apply a hierarchical class structure for FSL, benefiting from the good performance of it in hierarchical classification [25], [26]. A hierarchical structure is mainly constructed according to the multi-granularity semantic correlations in WordNet [14]. It showcases the correlations that exist between fine-grained and coarse-grained classes at multi-levels of granularity and can serve as a crucial external resource that guides the process of FSL [19], [20]. For the learning of feature representation, Zhu et al. [16] proposed multi-granularity episodic contrastive learning guided by the hierarchical structure. Several models employ the hierarchical structure to assist in FSL classifier construction. For instance, Li et al. [15] learned transferable visual features for large FSL classification by clustering a hierarchical structure. In addition, a coarse-to-fine FSL classifier is conducted by the hierarchical structure in [17]. Similarly, Zhang et al. [18] designed an interpretable decision tree-based classifier.

Although the above models have achieved promising success, they cannot make a connection between the class structural knowledge and limited visual information. For few-shot learning, the learning of limited visual information plays a decisive role in model performance. In this paper, our objective is to establish a strong connection and synergy between visual information and structural knowledge of classes to enhance FSL. Guided by the class structured knowledge, we integrate global and local visual information to explore intra-image and inter-class relationships, which enables models to learn more distinguishing class information. By integrating visual information and structural class knowledge to guide classi-

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



fication decisions, it enables models effectively learn from limited labeled examples and achieve improved generalization accuracy for new classes. Unlike the "coarse to fine" classifier, the fusion strategy can alleviate the problem of inter-layer error transmission. Additionally, they assume that the misclassification errors encountered in objective learning are equal, only considering whether the prediction is correct or not. In practical applications, different classification results should have different classification risks. Thereby, guided by the structural knowledge, a hierarchical loss is introduced to minimize the degree of classification error, using different weights according to the hierarchical structure. In summary, we strive to simultaneously study the visual information and structural knowledge, working in a mutual way to enhance FSL.

## III. Proposed Model

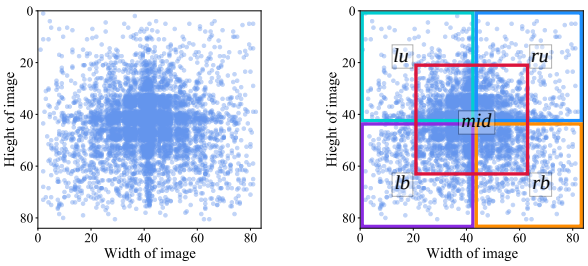
In this section, we give a brief description of MGKFD in detail, and the framework is shown in Fig. 1. MGKFD can be mainly divided into two parts: First, we divide the global features into five local features and obtain the multi-granularity features according to the hierarchical structure. Second, we fuse the multi-granularity knowledge and abundant visual information for similarity measurement and FSL classification decision-making. In addition, a hierarchical loss is established to minimize the classification error.

### A. Multi-granularity feature extraction

In this section, we obtain multi-granularity features combining the global and local features and the correlations of different granularity classes in the hierarchical class structure.

Many existing few-shot methods directly utilize the features extracted by the Convolutional Neural Network (CNN) for image similarity measurement and have achieved effective results. However, there are two disadvantages: 1) The features extracted by CNN represent the mixed global features of the image, which are fuzzy and not suitable for fine-grained classification. 2) The mixed global features degrade and lose the image local features, and it is difficult to distinguish the differences among images. We randomly extract 10,000 images from dataset *tieredimageNet* [27] and analyze the position distribution of the main object in the images, as shown in Fig 2a. The major objects are mainly distributed in the middle of images, and the rest is scattered around. Therefore, we re-extract the local features from the global feature, cropping uniformly the global feature into five regions including the left-upper ($lu$), right-upper ($ru$), middle ($mid$), left-bottom ($lb$), and right-bottom ($rb$) regions of the global feature, as followed in Fig. 2b. The length of the region is half of the global length.

For a few-shot task, a training episode strategy of "$N$-way $K$-shot" is adopted under the meta-learning framework, which is formed by support and query sets. The support set consists of $N$ classes with $K$ labelled samples selected randomly from a training set $\mathcal{D}_{train}$, and the remaining $K'$ samples of the $N$ classes constitute the query set. Initially, let the support set be $\mathcal{S} = \{(x_1, y_1), ..., (x_{n_s}, y_{n_s})\}$ and the query set be $\mathcal{Q} = \{(\tilde{x}_1, \tilde{y}_1), ..., (\tilde{x}_{n_q}, \tilde{y}_{n_q})\}$, where $x_{n_s}$ and $x_{n_q}$ are the sample numbers of the support and query sets, respectively. For a 1-shot ($K = 1$) setting, the feature of every support sample represents the feature of the class to which it belongs. We assume $\tilde{x}_i$ and $x_j$ to be any samples from the query set $\mathcal{Q}$ and the support set $\mathcal{S}$, respectively, and define the feature extraction CNN as $\mathcal{F}_\phi$, where $\phi$ is the parameter of the CNN. We feed $\tilde{x}_i$ and $x_j$ into $\mathcal{F}_\phi$ to obtain the global features




(a) Original distribution (b) Local region division

Fig. 2. Distribution of the center position of the objects in images. Parameters $lu, ru, mid, lb$, and $rb$ mean the left-upper, right-upper, middle, left-bottom, and right-bottom local regions, respectively.

$$\begin{cases} \tilde{\mathbf{X}}_i^{(g)} = \mathcal{F}_\phi(\tilde{x}_i) \\ \mathbf{X}_j^{(g)} = \mathcal{F}_\phi(x_j), \end{cases} \tag{1}$$

where $\tilde{\mathbf{X}}_i^{(g)}$ is the global feature of query sample $\tilde{x}_i$ and $\mathbf{X}_j^{(g)}$ is the global feature of the $j^{th}$ support class. For a $K$-shot ($K > 1$) setting, we use a structured full connected layer [9] to learn a better global class feature for all samples of each class.

Then, we crop the global features into five local regions to extract the local features as follows

$$\begin{cases} \tilde{\mathbf{X}}_i^{(l)} = [\tilde{\mathbf{L}}_i^{(lu)}; \tilde{\mathbf{L}}_i^{(ru)}; \tilde{\mathbf{L}}_i^{(mid)}; \tilde{\mathbf{L}}_i^{(lb)}; \tilde{\mathbf{L}}_i^{(rb)}] \\ \mathbf{X}_j^{(l)} = [\mathbf{L}_j^{(lu)}; \mathbf{L}_j^{(ru)}; \mathbf{L}_j^{(mid)}; \mathbf{L}_j^{(lb)}; \mathbf{L}_j^{(rb)}], \end{cases} \tag{2}$$

where $\tilde{\mathbf{X}}_i^{(l)}$ is the local feature of $\tilde{x}_i$, $\mathbf{X}_j^{(l)}$ means the local feature of the $j^{th}$ support class, $\tilde{\mathbf{L}}_i^{(lu)}$, $\tilde{\mathbf{L}}_i^{(ru)}$, $\tilde{\mathbf{L}}_i^{(mid)}$, $\tilde{\mathbf{L}}_i^{(lb)}$, and $\tilde{\mathbf{L}}_i^{(rb)}$ are the five local features of query sample $\tilde{x}$, and $\mathbf{L}_j^{(lu)}$, $\mathbf{L}_j^{(ru)}$, $\mathbf{L}_j^{(mid)}$, $\mathbf{L}_j^{(lb)}$, and $\mathbf{L}_j^{(rb)}$ mean the five local features of the $j^{th}$ support class. To sum up, we obtain the global-local features of $\tilde{x}_i$ and the $j^{th}$ support class

$$\begin{cases} \tilde{\mathbf{X}}_i = [\tilde{\mathbf{X}}_i^{(g)}; \tilde{\mathbf{X}}_i^{(l)}] \\ \mathbf{X}_j = [\mathbf{X}_j^{(g)}; \mathbf{X}_j^{(l)}]. \end{cases} \tag{3}$$

Visual representations can explicitly capture important characters and features that aid in understanding and recognizing objects or concepts. A global feature contains a crude representation of the image content, while local features can capture the images distinctive information. However, models are prone to sub-optimization and limited generalization due to the scarcity of training data. While structural knowledge refers to organized information about classes, typically represented in a hierarchical or relational manner. The structured knowledge

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



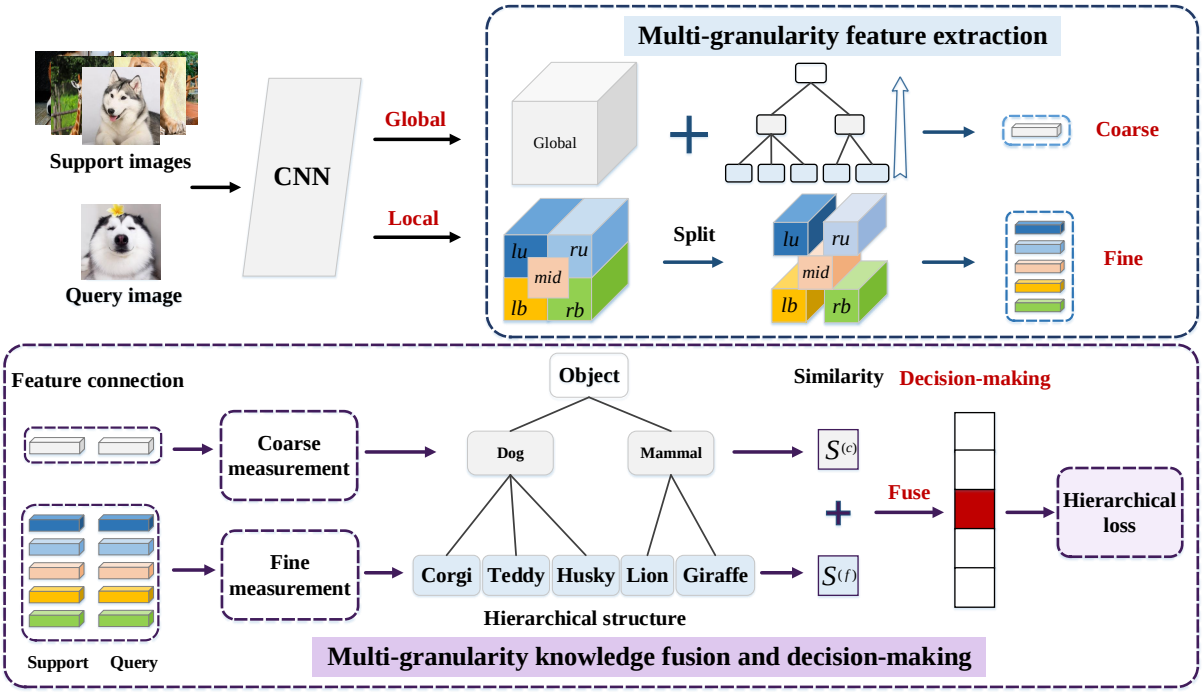



Fig. 1. Framework of MGKFD. MGKFD mainly consists of two parts: multi-granularity feature extraction and multi-granularity knowledge fusion and decision-making. Parameters $lu$, $ru$, $mid$, $lb$, and $rb$ mean the left-upper, right-upper, middle, left-bottom, and right-bottom region local features, respectively; $S^{(c)}$ and $S^{(f)}$ are the coarse- and fine-grained similarities, respectively.

like hierarchical class structure can guide the model in learning the similarities and variances among classes with few samples, enabling better generalization and discrimination capabilities for new tasks. In a hierarchical class structure, a coarse granularity class contains the commonness of the multiple fine-grained classes belonging to it and is relatively rough compared with fine-grained classes, while fine-grained classes reveal the variances among classes [20]. The hierarchical structure consists of coarse- and fine-grained classes constructed according to the multi-granularity correlations among classes [14]. The multi-granularity correlations are commonly obtained by measuring the semantic similarities of each two classes from WordNet.

Thereby, we leverage the hierarchical structure combined with the global and local features to explore more distinctive feature embedding. We represent the global features as coarse-grained class features and the local features as fine-grained class features. For a query sample, we define the global features as coarse-grained features and the local features as fine-grained features. For the support set, we define the local features as fine-grained features, and the coarse-grained class feature is calculated by the average of the global features of its fine-grained classes, as follows:

$$\mathbf{X}^{(c_i)} = \frac{1}{m_{c_i}} \sum_{k=1}^{m_{c_i}} \mathbf{X}_k^{(g)}, \tag{4}$$

where $c_i$ means the $i^{th}$ coarse-grained class of the support set and $m_{c_i}$ represents the number of its fine-grained classes.

Next, the following example shows the computing of the hierarchical features of the support samples.

***Example* 1:** In Fig. 3, there is a hierarchical structure with two coarse-grained and five fine-grained classes under a 5-way 1-shot setting. The features of each sample include global and local features. The hierarchical features of this example are calculated as follows: the fine-grained features are $\mathbf{X}^{(f_i)} = \mathbf{X}_i^{(l)}$, where $i = 1, ..., 5$; the coarse-grained features $\mathbf{X}^{(c_1)} = \frac{1}{3}(\mathbf{X}_1^{(g)} + \mathbf{X}_2^{(g)} + \mathbf{X}_3^{(g)})$ and $\mathbf{X}^{(c_2)} = \frac{1}{2}(\mathbf{X}_4^{(g)} + \mathbf{X}_5^{(g)})$.

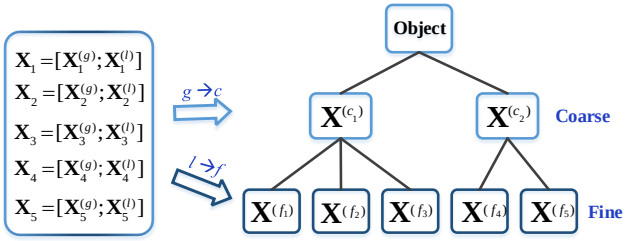



Fig. 3. An example of the hierarchical global-local feature extraction of the support samples. $g$: global feature, $l$: local feature, $c$: coarse-grained feature, and $f$: fine-grained feature.

## B. Multi-granularity knowledge fusion and decision-making

For the phase of classification, many few-shot models adopt the metric learning with the Cosine or Euclidean distances [7], [24]. Additionally, Zhang et al. [9] found that the Earth Mover's distance is better for exploiting the relevance of local features and can avoid the influence of the cluttered background. However, modelling by only visual information suffers from an over-fitting problem as available low-shot data.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



Therefore, by establishing a strong connection and synergy between visual information and structural knowledge of classes, we fuse the two knowledge to enhance the final decision-making of FSL, enabling models to learn more effectively from limited labeled examples and generalize to new classes.

Specially, we leverage the Earth Mover's distance to measure the similarities between the query sample and support multi-granularity features and fuse the multi-granularity similarities for final FSL decision-making. The Earth Mover's distance generates the optimal matching flows between the structural elements that have the minimum matching cost of the well-studied transportation problem, which is described as follows. A set of sources $\mathcal{S} = \{s_i | i = 1, ..., m\}$ transport goods to a set of destinations $\mathcal{D} = \{d_j | j = 1, ..., n\}$, where $s_i$ means the $i^{th}$ supply unit, $d_j$ represents the demand of the $j^{th}$ demander, and $m$ and $n$ are their total numbers, respectively. Let the cost per unit transported from the $i^{th}$ supplier to the $j^{th}$ demander be $c_{ij}$ and the number of units transported be $x_{ij}$. Then, the object is to find an optimal matching flow, as follows:

$$\min_{x_{ij}} \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}$$

subject to

$$\begin{aligned} & x_{ij} \geq 0,\ i = 1, ..., m, j = 1, ..., n \\ & \sum_{j=1}^{n} x_{ij} = s_i,\ i = 1, ..., m \\ & \sum_{i=1}^{m} x_{ij} = d_j,\ j = 1, ..., n, \end{aligned} \tag{5}$$

where $\mathcal{X} = \{x_{ij} | i = 1, ..., m, j = 1, ..., n\}$ means goods. We use the best matching cost between two feature vector sets to represent the similarity of two images.

We define $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_i, ..., \mathbf{u}_d]$ and $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_j, ..., \mathbf{v}_d]$ as two random features, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{H \times W \times C}$, $\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^{1 \times 1 \times C}$, $d = H \times W$, $C$ is the feature dimension, and $H$ and $W$ are height and width of feature map. Following the original earth mover's distance formulation in Eq. (5), the image similarity between two feature vector sets is calculated by

$$S(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{d}\sum_{j=1}^{d}(1 - c_{ij}), \tag{6}$$

where $d = H \times W$, $S(\cdot, \cdot)$ is applied to compute the similarity between feature vector sets and cost $c_{ij}$ represents the per unit matching between $\mathbf{u}_i$ and $\mathbf{v}_j$, as follows:

$$c_{ij} = 1 - \frac{{\mathbf{u}_i}^T \mathbf{v}_j}{\|\mathbf{u}_i\| \|\mathbf{v}_j\|}, \tag{7}$$

where similar vector sets create few matching costs between each other.

Furthermore, the similarities between different feature vectors are distinctive, which may be disadvantageous to the image similarity measurement. In Fig. 4, $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ are some feature subvectors of samples $u$ and $v$, respectively. Feature subvectors $\mathbf{u}_1$ and $\mathbf{v}_3$, $\mathbf{u}_3$ and $\mathbf{v}_2$

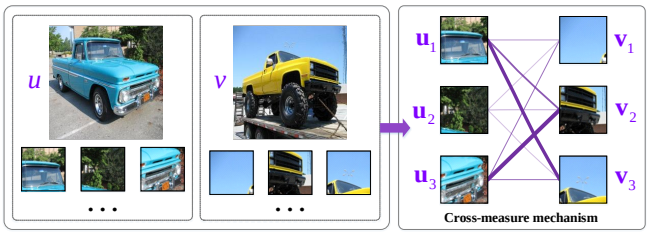



Fig. 4. An example for the visualization of feature matching. The thicker width of the line means a higher similarity between the two images.

are similar to each other, and the thicker and darker of their lines are. Local images $\mathbf{u}_3$ and $\mathbf{v}_1$ are less helpful for measuring image similarity, and the lines connecting them are not obvious. Then, we give different feature matching pairs different weights. The matching pair with large weights play a more important role in image matching, while that with small weights has little effect on the overall matching cost, regardless of which features are matched. The dot product is used to generate a weighted score between two feature vectors:

$$w_{ij} = max\{{\mathbf{u}_i}^T \mathbf{v}_j, 0\}, \tag{8}$$

where $w_{ij}$ denotes the weight between $\mathbf{u}_i$ and $\mathbf{v}_j$, and $max$ guarantees non-negative weights. In general, the weights are normalized as follows

$$\hat{w}_{ij} = \frac{w_{ij}}{\sum_{k=1}^{d} w_{kj}}, \tag{9}$$

where $\hat{w}_{ij}$ is the normalized weight, $H$ represents the height of the feature map, and $W$ is the width. Therefore, we obtain the weighted image similarity by combining Eqs. (6) and (9) as follows:

$$S(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{d}\sum_{j=1}^{d} \hat{w}_{ij}(1 - c_{ij}). \tag{10}$$

In the image similarity measurement, a higher image similarity has a smaller matching cost. In contrast, the greater the weight, the more favorable the similarity measurement of the image.

Then, we fuse the multi-granularity similarities to obtain the hierarchical similarity, which uses the knowledge of the coarse-grained class to guide fine-grained classification. The hierarchical similarity is calculated as

$$S_{ij}^{(h)} = \lambda S_{ij}^{(c)} + (1 - \lambda) S_{ij}^{(f)}, \tag{11}$$

where $c$ means the coarse-grained class, $f$ represents the fine-grained class, and $S_{ij}^{(h)}$ is the hierarchical similarity between query sample $\tilde{x}_i$ and the $j^{th}$ support class, and $\lambda$ is a balance factor. Similarities $S_{ij}^{(c)}$ and $S_{ij}^{(f)}$ denote the coarse- and fine-grained similarities, respectively, which are computed as:

$$\begin{cases} S_{ij}^{(c)} = S(\tilde{\mathbf{X}}^{(c_i)}, \mathbf{X}^{(c_j)}) \\ S_{ij}^{(f)} = S(\tilde{\mathbf{X}}^{(f_i)}, \mathbf{X}^{(f_j)}), \end{cases} \tag{12}$$

where $\tilde{\mathbf{X}}^{(c_i)}$ and $\tilde{\mathbf{X}}^{(f_i)}$ are the coarse- and fine-grained features of query sample $\tilde{x}_i$; $\mathbf{X}^{(f_j)}$ is the fine-grained feature of the $j^{th}$ support class and $\mathbf{X}^{(c_j)}$ is the coarse-grained feature to

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



which it belongs. In this way, by fusing the multi-granularity knowledge, we make the final decision according to the hierarchical similarity obtained, and the support fine-grained class with the highest hierarchical similarity is the predicted class of query samples.

We leverage the cross-entropy loss to compute the classification loss. The classification loss is computed as

$$\mathcal{L} = -\frac{1}{n_q}\sum_{i=1}^{n_q}\sum_{j=1}^{N} y_{ij} log(\frac{exp(-S_{ij}^{(h)})}{\sum_{k=1}^{N} exp(-S_{ik}^{(h)})}), \tag{13}$$

where $y_{ij} = 1$ if the real class of query sample $\tilde{x}_i$ is the $j^{th}$ class and $y_{ij} = 0$ otherwise; $S_{ij}^{(h)}$ is the hierarchical similarity between the query sample $\tilde{x}_i$ and the $j^{th}$ support class; $n_q$ means the number of all query samples, and $N$ denotes the number of classes in every $N$-way $K$-shot episode.

However, the standard cross-entropy loss function equally treats all classification errors with the *one-hot* strategy, which defines that there are only two cases of prediction results: correct or incorrect. For instance, the classification errors are the same whether the query sample is incorrectly classified as *Bear* or *Butterfly* in Fig. 5(a). In fact, the classification errors are distinctive. Fig. 5(b) gives an intuitive explanation. Compared with class *Butterfly*, class *Bear* is closer to class *Panda* according to the hierarchical structure of the support classes, because *Bear* and *Panda* both belong to the coarse-grained class *Mammal*, while *Butterfly* belongs to *Insect*. From the relations among classes, the degree of classification error for the query sample classified as *Butterfly* is higher than that classified as *Bear*.

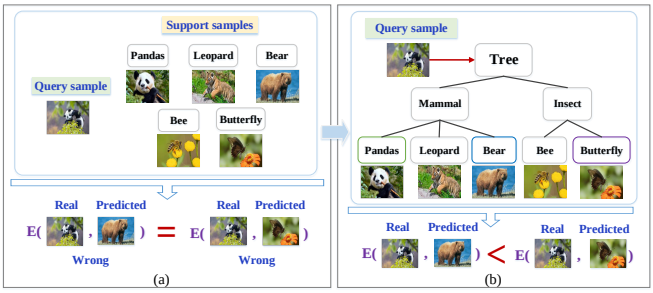



Fig. 5. Different classification errors. E$(\cdot,\cdot)$ means the degree of classification error.

Therefore, we design a hierarchical loss by using the hierarchical structure to weigh the classification results differently. Different classification errors result from different levels of penalty. We introduce the *tree induced error* (*TIE*) to weigh the classification error risks. *TIE* can reflect the sample misclassification degree by measuring the number of edges passed from the true class to the predicted class in the hierarchical structure. Its formal representation is as follows:

$$TIE(\tilde{y}_i, y_j) = |E_H(\tilde{y}_i, y_j)|, \tag{14}$$

where $\tilde{y}_i$ is the true class of query sample $\tilde{x}_i$, and $y_j$ is its predicted class; $E_H(\tilde{y}_i, y_j)$ represents the edge set of the path from $\tilde{y}_i$ to $y_j$ and $|\cdot|$ is the number of elements in a set. Then, we define the weight of classification error as

$$t_{ij} = \frac{TIE(\tilde{y}_i, y_j)}{max\ TIE(\tilde{y}_i, *)}, \tag{15}$$

where $t_{ij}$ is the weight of classification error of query sample $\tilde{x}_i$ when it is classified as the $j^{th}$ class, and $max\ TIE(\tilde{y}_i, *)$ denotes the max *TIE* between the predicted class and the real class $\tilde{y}_i$ of $\tilde{x}_i$. Next, we use the ***Example* 2** to show a computing example of *TIE*.

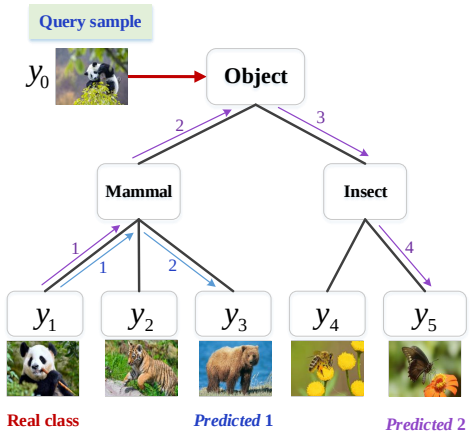



Fig. 6. An example for the weight computing of classification error risks. Arrows of different colors represent paths between different classes.

***Example* 2:** In Fig. 6, *Predicted* 1 and *Predicted* 2 are two classification results of the query sample. In Fig. 6, $TIE(y_0, y_1) = 0$, $TIE(y_0, y_3) = 2$, $TIE(y_0, y_5) = 4$ and $max\ TIE(y_0, *) = 4$, then the weights are $t_{01} = \frac{0}{4} = 0$, $t_{03} = \frac{2}{4} = \frac{1}{2}$ and $t_{05} = \frac{4}{4} = 1$. $TIE(y_0, y_1) = 0$ indicates that the classification is correct. $t_{03} < t_{05}$ denotes the degree of classification error for the query sample classified as *Predicted* 1, which is less than that classified as *Predicted* 2.

Then, we obtain the hierarchical loss $\mathcal{L}_h$ by combining Eqs. (13) and (15) as follows:

$$\mathcal{L}_h = -\frac{1}{n_q}\sum_{i=1}^{n_q}\sum_{j=1}^{N}(1+t_{ij}) y_{ij} log(\frac{exp(-S_{ij}^{(h)})}{\sum_{k=1}^{N} exp(-S_{ik}^{(h)})}), \tag{16}$$

where $y_{ij} = 1$ if the real class of query sample $\tilde{x}_i$ is the $j^{th}$ class, and $y_{ij} = 0$ otherwise. The hierarchical loss degenerates to the standard cross entropy loss when the weights are 0.

Algorithm 1 describes the MGKFD training process. In particular, before training, we obtain hierarchical class structures of different datasets by WordNet (We use hierarchical structures constructed from datasets that are publicly accessible). According to the hierarchical class structure, we define the all classes as fine-grained classes and obtain their coarse-grained classes, and then extract multi-granularity features of query and support sets in lines 5-10. Then, we fuse the multi-granularity knowledge to obtain hierarchical similarities and make the final classification decision in lines 12-14. Next, we compute the weights of classification error by *TIE* according

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



to the structure and update the hierarchical loss in lines 15-17. Finally, we backpropagate the hierarchical loss to update the parameter $\phi$.

**Algorithm 1** Few-shot Learning with Multi-granularity Knowledge Fusion and Decision-making (MGKFD)

**Input**: Training set $\mathcal{D}_{train}$. Setting $E_{max}$ as the total training epochs. The feature extraction module is $\mathcal{F}_\phi$.
**Output**: The parameter $\phi$.

1: Initialize the parameter $\phi$;
2: Obtain the hierarchical structure $T$ of the classes in $\mathcal{D}_{train}$ by WordNet;
3: Define the all classes as fine-grained classes and obtain their coarse-grained classes according to the structure $T$;
4: **for** $epo = 1 : E_{max}$ **do**
5: Construct a task consisting of the query set $\mathcal{Q}$ and support set $\mathcal{S}$ respectively;
6: Extract the global features of $\mathcal{Q}$ and $\mathcal{S}$ by $\mathcal{F}_\phi$;
7: Obtain the local features of $\mathcal{Q}$ and $\mathcal{S}$;
8: Obtain the fine- and coarse-grained (local and global) features of $\mathcal{Q}$ and the fine-grained features of $\mathcal{S}$;
9: Obtain the coarse-grained classes in the task according to the structure $T$;
10: Compute the coarse-grained features of $\mathcal{S}$ by Eq. (4);
11: **for** $i = 1 : n_q$ **do**
12: Compute the multi-granularity similarities by Eq. (12);
13: Obtain the hierarchical similarity by Eq. (11);
14: Make the decision of $\tilde{x}_i$ according to the max hierarchical similarity;
15: Compute the *TIE* of the predicted class and truth class of $\tilde{x}_i$ by Eq. (14);
16: Compute the weight of classification error of $\tilde{x}_i$ by Eq. (15);
17: Update the hierarchical loss by Eq. (16);
18: **end for**
19: Update the parameter $\phi$ by the hierarchical loss back-propagation;
20: **end for**

## IV. Experimental Settings

### A. Implementation details

In the experiments, we use ResNet12 as our network backbone and SGD as the optimizer, following recent few-shot classification work [9], [24]. As is commonly implemented in the advanced literature, we use a feature pre-training step to train our network followed [9]. We set the region length as half of the global feature length. The values of $\lambda$ are 0.4 on *tieredImageNet* and 0.6 on *FC100* and *CIFAR-FS*. All images are resized to $84 \times 84 \times 3$. For the test, we randomly select 5,000 episode samples from the test set with a 95% confidence interval. The experiments are conducted using a GeForce RTX 2080 Ti Nvidia GPU and Pytorch. The code and datasets used in the proposed model will be opened to GitHub (https://github.com/fhqxa/MGKFD).

### B. Datasets

For the FSL experiments, we use three popular benchmark datasets: *tieredImageNet* [27], *FC100* [29] and *CIFAR-FS* [30]. The basic statistics for these datasets are as follows:

***tieredImageNet*** [27] is a few-shot sub-dataset originating from the ILSVRC-12 dataset, and includes 779,765 images in total. There are 608 fine-grained classes and 34 coarse-grained classes. The training/validation/test set splits consist of 20/6/8 coarse-grained classes, which are supersets of 351/97/160 fine-grained classes.

***FC100*** [29] is a popular dataset for few-shot classification, which is built from *CIFAR100*. There are 100 fine-grained classes and 20 coarse-grained classes. The respective training/validation/test set splits consist of 12/4/4 coarse-grained classes, which are supersets of 60/20/20 fine-grained classes. The hierarchical structure is shown in Fig. 7.

***CIFAR-FS*** [30] also originate from the *CIFAR100* dataset as *FC100*. There are 100 fine-grained classes and 20 coarse-grained classes. The respective training/validation/test set splits consist of 20/11/13 coarse-grained classes, consisting of 64/16/20 fine-grained classes. Their coarse-grained classes are shared.

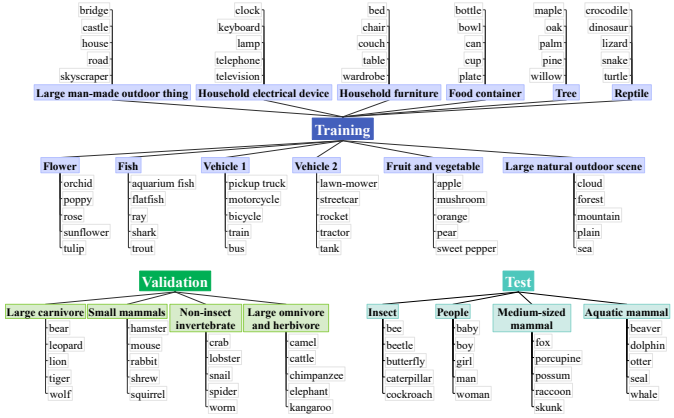



Fig. 7. Hierarchical structure of *FC100* with the training, validation, and test sets.

### C. Evaluation measures

We adopt five evaluation measures to verify the effectiveness of our model: Classification Accuracy (*ACC*), Area Under Curve (*AUC*), F1 score ($F_1$), Tree Induced Error (*TIE*), and Hierarchical $F_1$ score ($F_H$). *TIE* and $F_H$ are used for hierarchical classification model evaluations. $F_H$ is a joint calculation of hierarchical precision ($P_H$) and hierarchical recall ($R_H$):

$$P_H = \frac{|\hat{D}_{aug} \cap D_{aug}|}{|\hat{D}_{aug}|}, \ R_H = \frac{|\hat{D}_{aug} \cap D_{aug}|}{|D_{aug}|},$$

where $\hat{D}_{aug}$ contains the predicted class and its ancestor nodes, while $D_{aug}$ represents the true class and its ancestor nodes. The $F_H$ is calculated as follows

$$F_H = \frac{2P_H R_H}{P_H + R_H}.$$

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



## V. Experimental results and analysis

In this section, we conduct several ablation studies, comparison experiments, and visualization to verify and analyse the effectiveness of MGKFD. For the convenience of writing and understanding, we set some proper nouns into abbreviations: global and local features (*g-l*), multi-granularity knowledge fusion and decision-making (*KFD*), hierarchical loss (*HL*), Mean Square Error loss (*MSE*), and Cross-Entropy loss (*CE*).

### A. Ablation studies

**Verification of multi-granularity knowledge fusion and decision-making strategy.** By establishing a strong connection and synergy between global and local visual information and multi-grained class knowledge, we fuse the two knowledge to enhance the final decision-making of FSL. We construct ablation studies to validate every component on *tieredImageNet* and *FC100*. The performance of the different components is presented in Fig. 8. We can find that: Each component is beneficial to the multi-granularity strategy. By the division of global and local features, we obtain 71.94/44.33% $ACC$, which are about 1.20/0.70% better than the baseline on the two datasets, respectively. Likewise, *KFD* makes about 1.00/0.40% improvements over the baseline. Moreover, by integrating both strategies, our model has more improvement. The experiments validate that the combining of global and local information and multi-granularity knowledge is beneficial for the classification of fine-grained classes.

To intuitively represent the effectiveness of *KFD*, we visualize the different inference results of *KFD* on *tieredImageNet*, and the classification visualization is shown in Fig. 9. We can find that the final predictions of fine granularity classes are corrected with the assistance of the coarse granularity classes. For example, the sample with the real class of *Beer bottle* originally wrongly divided into class *Eggnog* is correctly classified into *Beer bottle* with the influence of coarse granularity class. It proves that the class multi-granularity correlations play a positive and guiding role in the classification ability of MGKFD.

**Verification of hierarchical loss.** We validate the effect of HL compared with other loss functions, including the mean square error and cross-entropy. We use three evaluation measures to estimate the performance of the model. The performance is

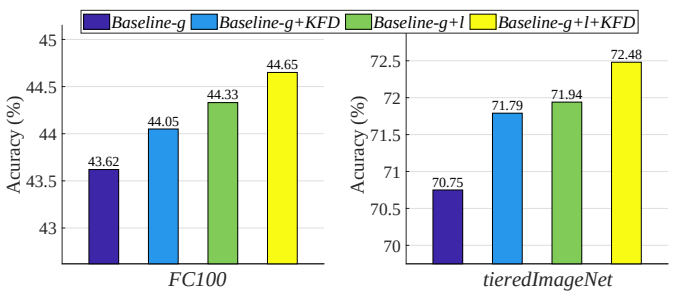



Fig. 8. $ACC$ (%) of different components of *KFD* on *tieredImageNet* and *FC100* (5-way 1-shot).

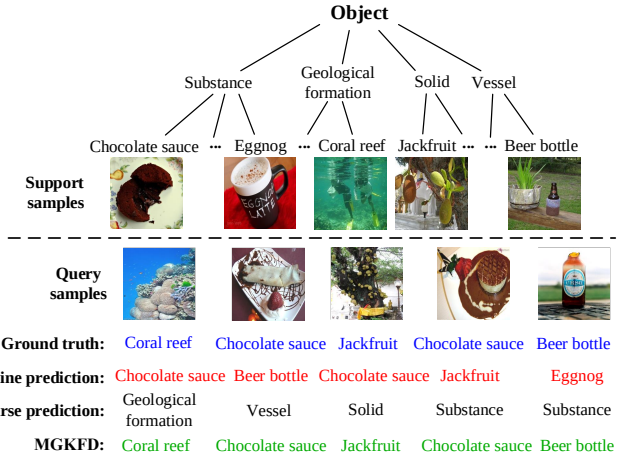



Fig. 9. Classification visualization of MGKFD on *tieredImageNet* dataset.

TABLE I
$ACC$ (%) OF MGKFD WITH DIFFERENT LOSS FUNCTIONS ON DIFFERENT DATASETS (5-WAY 1-SHOT). THE BEST RESULTS ARE MARKED IN BOLD. $ACC$ (↑), *TIE* (↓) AND $F_H$ (↑). "↑" REPRESENTS THAT THE BIGGER THE VALUE, THE BETTER THE PERFORMANCE. "↓" INDICATES THAT THE SMALLER THE BETTER.

| Model | Loss | *tieredImageNet* | | | *FC100* | | |
|---|---|---|---|---|---|---|---|
| | | $ACC$ | *TIE* | $F_H$ | $ACC$ | *TIE* | $F_H$ |
| MGKFD | *MSE* | 72.26 | 63.87 | 85.16 | 43.92 | 78.04 | 68.90 |
| MGKFD | *CE* | 72.48 | 63.76 | 85.38 | 44.65 | 77.67 | 69.56 |
| MGKFD | *HL* | **72.78** | **63.61** | **85.53** | **44.86** | **77.57** | **69.68** |

compared in Table I, and the main results are: 1) We observe a gain in performance when adopting *HL*, with better gains than when using *MSE* and *CE*. Specifically, with *HL*, the $ACC$s are about 0.50% and 0.90% better than with *MSE* on *tieredImageNet* and *FC100* respectively, similar to $F_H$. This indicates that the hierarchical loss promotes the classification ability of the model. 2) *HL* can make MGKFD with low misjudgment degree of misjudgment. The *TIE* value reflects the misclassification degree of the model. The *TIE* of MGKFD with *HL* is less than that with *MSE* and *CE*, highlighting the benefits of applying *TIE* to minimize the classification loss with different weights.

**Verification of different components proposed.** We describes an ablation study that is conducted to validate each component of the MGKFD, including the division of global and local features, multi-granularity knowledge fusion and decision-making, and hierarchical loss. The main experimental results obtained on *tieredImageNet* and *FC100* are listed in Table II, and the following conclusions can be made: 1) The results illustrate the effectiveness of using these key components in our model. The effect of *KFD+g-l* is greater than that of *HL* in MGKFD. Compared with the baseline, improvements of 0.60% and 0.20% are obtained by using *HL*, respectively. Moreover, we obtain even better improvements of 1.70% and 1.00% by *KFD+g-l* on the two datasets, respectively. 2) *KFD* and *HL* tend to achieve greater improvements on *tieredImageNet* than on *FC100*. One possible reason is that the effects of *KFD* and *HL* are related to the class hierarchy

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



TABLE II
CONTRIBUTIONS ($ACC$ %) OF DIFFERENT COMPONENTS IN MGKFD ON DIFFERENT DATASETS (5-WAY 1-SHOT). BEST RESULTS ARE MARKED IN BOLD.

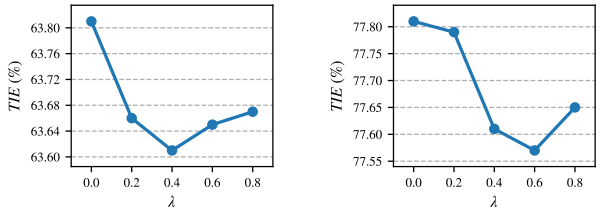
| *g-l* | *KFD* | *HL* | *tieredImageNet* | *FC100* |
|---|---|---|---|---|
| | | | 70.75 | 43.62 |
| | | ✓ | 71.34 | 43.81 |
| | ✓ | | 71.79 | 44.05 |
| ✓ | | | 71.94 | 44.33 |
| ✓ | | ✓ | 72.33 | 44.35 |
| ✓ | ✓ | | 72.48 | 44.65 |
| ✓ | ✓ | ✓ | **72.78** | **44.86** |

tree of the data. The class scale of *tieredImageNet* is larger than that of *FC100*, which presents richer class information that facilitates model training. 3) Using the three components together, MGKFD achieves 72.48% and 44.86% $ACC$s on the two datasets, outperforming the baseline by about 2.00% and 1.20%. The MGKFD is most effective when using all of the components. The consistent improvements across the two datasets empirically validate our claim that joint optimization of the hierarchical structural knowledge and global and local visual information is beneficial for FSL.

**Parameter analysis.** We adopt a weighted aggregation technique of $\lambda$ to fuse the multi-grained knowledge and global and local visual information to make classification decisions as Eq. (11). The variation in the parameter represents the degree of fusion between coarse- and fine-grained knowledge. We use three evaluation measures for the experiments, and the performance of different $\lambda$ values are shown in Table III and Fig. 10. We can conclude that: Larger values of $ACC$ and $F_H$ result in better model performance, while the *TIE* value is smaller and the models degree of misclassification is lower. We note that the best performance is achieved by setting $\lambda = 0.4, 0.6$ on *tieredImageNet* and *FC100*, respectively. This indicates that with an appropriate $\lambda$ value, the coarse-grained class representations and discriminative knowledge can be helpful for the learning of fine granularity since they have strong semantic relations with fine-grained classes.

TABLE III
$ACC$ AND $F_H$ (%, ↑) OF DIFFERENT $\lambda$ VALUES ON DIFFERENT DATASETS (5-WAY 1-SHOT). "↑" REPRESENTS THAT THE BIGGER THE VALUE, THE BETTER THE PERFORMANCE. THE BEST RESULTS ARE MARKED IN BOLD.

| $\lambda$ | *tieredImageNet* | | *FC100* | |
|---|---|---|---|---|
| | $ACC$ | $F_H$ | $ACC$ | $F_H$ |
| 0 | 72.37 | 85.25 | 44.38 | 69.20 |
| 0.2 | 72.67 | 85.45 | 44.41 | 69.39 |
| 0.4 | **72.78** | **85.53** | 44.79 | 69.65 |
| 0.6 | 72.70 | 85.47 | **44.86** | **69.68** |
| 0.8 | 72.66 | 85.45 | 44.70 | 69.54 |




(a) *tieredImageNet* (b) *FC100*

Fig. 10. *TIE*(↓) of different $\lambda$ values on different datasets. "↓" indicates that the smaller the better.

## B. Performance comparison

A comparison between MGKFD and several advanced FSL models is conducted. We select comparison models with four types: 1) *single-scale-global*: modelling by global visual features; 2) *single-scale-local*: modelling by local visual features; 3) *multi-scale*: modelling by global and local visual features; 4) *multi-granularity*: modelling by multi-grained knowledge of the hierarchical class structure. The first three methods strive to mining global and local information to improve model performance, and the last one mostly focuses on embedding the class structural knowledge neglecting the learning of visual information. While MGKFD aims to simultaneously explore the both knowledge, working in a mutual way to enhance the performance of FSL.

The results on three datasets, including *tieredImageNet*, *FC100*, and *CIFAR-FS*, are shown in Tables IV and V. We can obtain the following observations: 1) MGKFD outperforms most single- and multi-scale models. More specifically, compared with the single-scale method, multi-scale and multi-granularity methods have a natural advantage for limited data, as they integrate more information. MGKFD achieves 72.78% $ACC$ in 1-shot setting, which is about 3.00% and 1.30% better than RFS-simple and infoPatch on *tieredImageNet*, respectively. Unlike infoPatch which randomly blocks part of the images to focus on local objects, we re-extract the global feature into five local features and cross-measure them to reduce the interference from background knowledge. Compared with the high- and low-level multi-scale features

TABLE IV
ACCURACY COMPARISON WITH DIFFERENT MODELS ON *tieredImagenet* ($ACC$%). THE BEST RESULTS ARE MARKED IN BOLD.

| **Model** | **Method** | **5-way 1-shot** | **5-way 5-shot** |
|---|---|---|---|
| MatchNet [6] | *single-scale-global* | $68.50 \pm 0.92$ | $80.60 \pm 0.71$ |
| ProtoNet [7] | | $68.23 \pm 0.23$ | $84.03 \pm 0.16$ |
| MetaOptNet [23] | | $65.99 \pm 0.72$ | $81.56 \pm 0.53$ |
| CTM [31] | | $68.41 \pm 0.39$ | $84.28 \pm 1.73$ |
| RFS-simple [32] | | $69.74 \pm 0.72$ | $84.41 \pm 0.55$ |
| MeTAL [33] | | $63.89 \pm 0.43$ | $80.14 \pm 0.40$ |
| DeepEMD [9] | *single-scale-local* | $71.71 \pm 0.31$ | $85.86 \pm 0.21$ |
| LMPNet [34] | | $70.21 \pm 0.15$ | $79.45 \pm 0.17$ |
| infoPatch [35] | | $71.51 \pm 0.52$ | $85.44 \pm 0.35$ |
| RENet [24] | | $70.79 \pm 0.50$ | $84.93 \pm 0.35$ |
| QPN [36] | *multi-scale* | $55.96 \pm 0.77$ | $75.01 \pm 0.84$ |
| MSML [10] | | $70.24 \pm 1.99$ | $83.03 \pm 1.42$ |
| AMTIP [12] | | 60.99 | 75.33 |
| wDAE-GNN [37] | *multi-granularity* | $68.18 \pm 0.16$ | $83.09 \pm 0.12$ |
| HGNN [25] | | $64.32 \pm 0.49$ | $83.34 \pm 0.45$ |
| CGFSC [38] | | $60.54 \pm 0.79$ | $75.22 \pm 0.63$ |
| HMRN [39] | | $57.98 \pm 0.26$ | $74.70 \pm 0.24$ |
| MGECL [16] | | $71.48 \pm 0.40$ | $\mathbf{86.17 \pm 0.27}$ |
| MGKFD (ours) | | $\mathbf{72.78 \pm 0.31}$ | $85.86 \pm 0.22$ |

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2024.3350542



TABLE V
ACCURACY COMPARISON WITH DIFFERENT FLAT MODELS ON *FC100* AND *CIFAR-FS* (*ACC%*). THE BEST RESULTS ARE MARKED IN BOLD.

**(a) *FC100***

| Model | Method | 5-way 1-shot | 5-way 5-shot |
|---|---|---|---|
| ProtoNet [7] | *single-scale-global* | $41.54 \pm 0.76$ | $57.08 \pm 0.76$ |
| MetaOptNet [23] | | $41.10 \pm 0.60$ | $55.50 \pm 0.60$ |
| BFS [40] | | $43.16 \pm 0.59$ | $57.57 \pm 0.55$ |
| RFS-simple [32] | | $42.60 \pm 0.70$ | $59.10 \pm 0.60$ |
| ConstellationNet [41] | *single-scale-local* | $43.80 \pm 0.20$ | $59.70 \pm 0.20$ |
| DeepEMD [9] | | $43.44 \pm 0.24$ | $60.80 \pm 0.25$ |
| infoPatch [35] | | $43.80 \pm 0.40$ | $58.00 \pm 0.40$ |
| RENet [24] | | $43.63 \pm 0.41$ | $59.89 \pm 0.40$ |
| SSFormers [42] | *multi-scale* | $43.72 \pm 0.21$ | $58.92 \pm 0.18$ |
| CGFSC [38] | *multi-granularity* | $39.57 \pm 0.23$ | $54.63 \pm 0.63$ |
| MGKFD (ours) | | $\mathbf{44.86 \pm 0.26}$ | $\mathbf{61.10 \pm 0.25}$ |

**(b) *CIFAR-FS***

| Model | Method | 5-way 1-shot | 5-way 5-shot |
|---|---|---|---|
| ProtoNet [7] | *single-scale-global* | $55.5 \pm 0.70$ | $72.0 \pm 0.60$ |
| Shot-Free [43] | | 69.15 | 84.70 |
| Cosine Classifier [44] | | $69.39 \pm 0.28$ | $72.85 \pm 0.65$ |
| S2M2 [45] | | $62.77 \pm 0.23$ | $75.75 \pm 0.13$ |
| MeTAL [33] | | $67.97 \pm 0.47$ | $82.17 \pm 0.38$ |
| RCN [46] | *single-scale-local* | $69.02 \pm 0.92$ | $82.96 \pm 0.67$ |
| TEAM [47] | *multi-scale* | 71.43 | 81.25 |
| HGNN [25] | *multi-granularity* | $60.44 \pm 0.51$ | $81.05 \pm 0.44$ |
| MGKFD (ours) | | $\mathbf{72.96 \pm 0.30}$ | $\mathbf{86.12 \pm 0.21}$ |

of MSML, our multi-granularity global-local features show greater advantages. 2) Under the *multi-granularity* strategy, MGKFD shows advantages over most models. wDAE-GNN obtains good performance by adopting a graph structure to construct class correlations, while our model obtains better gains using the multi-granularity class relations in the hierarchical structure. Moreover, the classification ability of our model is superior to those of HMRN and HGNN, despite them employing the same hierarchical class structure as auxiliary knowledge. But MGKFD is slightly inferior to MGECL in 5-shot setting. The reason for our reasoning may be that MGECL adopts contrastive learning to learn distinguishing features, which is more applicable when the samples are not very few. In summary, MGKFD enables models effectively learn from limited labeled examples and achieve improved generalization accuracy for new classes by fusing the visual information and structural knowledge to make classification decisions.

For the computational complexity, our model adopts a common convolutional network architecture, which is used in most of the existing models such as [16], [24], [35]. Based on [9], our main work is in the phase of inference and train the architecture by a specific loss function proposed without adding any network. Calculate by formula, the times of the multi-grained knowledge embedding is almost negligible. Therefore, our model shares the same algorithmic complexity with these existing models. Additionally, it allows us transferring our model to other models where the framework like [9], such as [7], [24] since our main contributions are in the inference step and the design of loss.

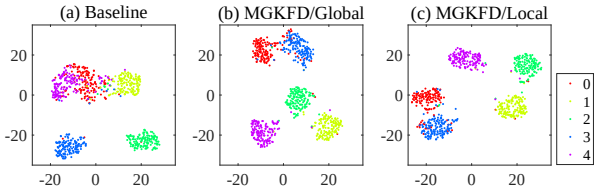



Fig. 11. t-SNE visualization on *tieredImageNet* (5-way 1-shot). Numbers 0-4 mean different fine-grained classes. Classes 0 and 3 belong to the same coarse-grained class, and other classes belong to an individual coarse-grained class.

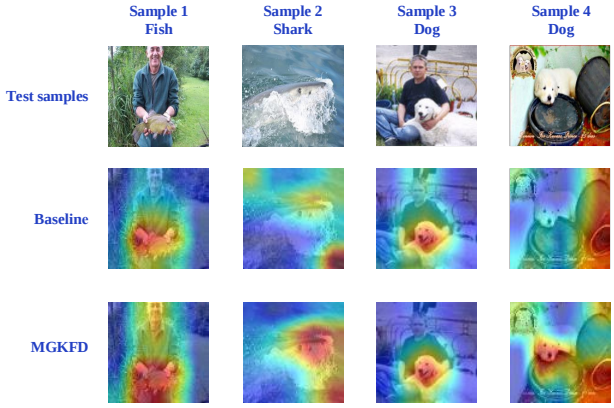



Fig. 12. Heatmap visual comparison between MGKFD and baseline.

## C. Visualization

We adopt T-SNE technique to display the distribution of test visual features before and after trained by MGKFD, as shown in Fig. 11. We can find that 1) the inter-class boundary is clearer when using global features, and the classes that belong to the same coarse-grained class are more compact than that of the baseline. 2) The classification boundary of each class is more obvious, and the intra-class distance is reduced with the local features in Fig. 11 (c). This indicates that the combination of global and local features and the multi-granularity classes can obtain more distinctive feature embedding.

In addition, we use a heatmap to visualize the spatial correspondence of MGKFD, with the main results shown in

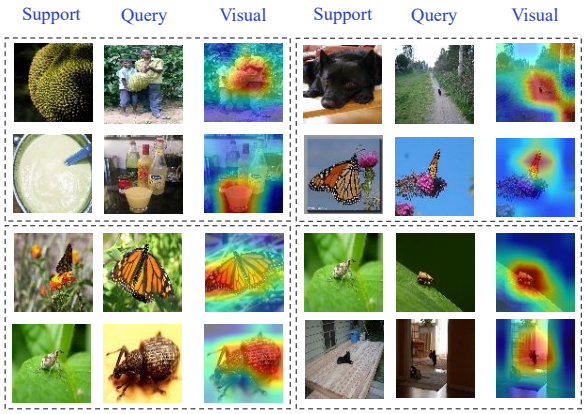



Fig. 13. Heatmap visualization of MGKFD under different scenarios.

Figs. 12 and 13. We can observe that our model outperforms the baseline in terms of spatial relationships. The MGKFD covers objects more accurately by weakening the influence of dominant backgrounds under different scenarios.

## VI. Conclusions and Future Work

In this paper, we present a unified FSL framework based on multi-granularity knowledge fusion and decision-making (MGKFD). It considers the rich visual information and strong class structural knowledge simultaneously, working in a mutual way to enhance FSL. Multi-granularity knowledge is employed to obtain more distinguishing class features by combining the global and local features, which more effectively mines information from limited data. The multi-granularity knowledge fusion and decision-making are adopted to fuse the multi-granularity knowledge to make the final few-shot classification decision. Since the representations and discriminative knowledge of coarse-grained can be beneficial for the learning of fine granularity. In addition, we design a hierarchical loss to weigh the different classification results to minimize the classification loss, with consideration of the correlations between the predicted and real classes. We leverage three evaluation measures to validate the effectiveness of MGKFD on three popular few-shot datasets. The experimental results show that MGKFD is comparable with several advanced few-shot learning models.

There are some limitations of MGKFD which will be the future study. In the step of multi-grained knowledge fusion, it is time-consuming to adjust the weights of different granularities on different datasets. We will focus on designing an adaptive fusion strategy to choose optimal weights. In addition, the hierarchical structure is a relatively chivalrous body of knowledge, with limited generalization ability. While knowledge graph encompasses a broader range of objects and domains, possessing a wealth of knowledge and enhanced generalization capabilities. The rich connections and information in knowledge graph will be further mined to guide few-shot learning in our future work.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62141602 and 62076116, the Natural Science Foundation of Fujian Province under Grant Nos. 2021J011003 and 2021J011004, and the Ministry of Education Industry-University-Research Innovation Program under Grant No. 2021LDA09003.

## References

[1] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *CCSS*, vol. 33, 2011, pp. 2568–2573.

[2] E. Miller, N. Matsakis, and P. Viola, "Learning from one example through shared densities on transforms," in *TIPAMI*, vol. 1, 2000, pp. 464–471.

[3] M. Jamal and G. Qi, "Task agnostic meta-learning for few-shot learning," in *CVPR*, 2019, pp. 11 719–11 727.

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

[5] Y. Wang, Q. Yao, J. Kwok, and L. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.

[6] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016, pp. 3637–3645.

[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017, pp. 4080–4090.

[8] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. Torr, and T. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018, pp. 1199–1208.

[9] C. Zhang, Y. Cai, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *CVPR*, 2020, pp. 12 203–12 213.

[10] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *TCSVT*, vol. 31, no. 3, pp. 1091–1102, 2020.

[11] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, and J. Xue, "Bsnet: Bi-similarity network for few-shot fine-grained image classification," *TIP*, vol. 30, pp. 1318–1331, 2021.

[12] S. Fu, B. Liu, W. Liu, B. Zou, X. You, Q. Peng, and X.-Y. Jing, "Adaptive multi-scale transductive information propagation for few-shot learning," *KBS*, vol. 249, p. 108979, 2022.

[13] H. Chen, R. Liu, Z. Xie, Q. Hu, J. Dai, and J. Zhai, "Majorities help minorities: Hierarchical structure guided transfer learning for few-shot fault recognition," *PR*, vol. 123, p. 108383, 2022.

[14] H. Zhao, Q. Hu, P. Zhu, Y. Wang, and P. Wang, "A recursive regularization based feature selection framework for hierarchical classification," *TKDE*, vol. 33, no. 7, pp. 2833–2846, 2021.

[15] A. Li, T. Luo, Z. Lu, T. Xiang, and L. Wang, "Large-scale few-shot learning: Knowledge transfer with class hierarchy," in *CVPR*, 2019, pp. 7212–7220.

[16] P. Zhu, Z. Zhu, Y. Wang, J. Zhang, and S. Zhao, "Multi-granularity episodic contrastive learning for few-shot learning," *PR*, vol. 131, p. 108820, 2022.

[17] L. Liu, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Many-class few-shot learning on multi-granularity class hierarchy," *TKDE*, vol. 34, no. 05, pp. 2293–2305, 2022.

[18] B. Zhang, H. Jiang, X. Li, S. Feng, Y. Ye, C. Luo, and R. Ye, "Metadt: Meta decision tree with class hierarchy for interpretable few-shot learning," *TCSVT*, 2023.

[19] S. Wang, X. Chen, Y. Wang, M. Long, and J. Wang, "Progressive adversarial networks for fine-grained domain adaptation," in *CVPR*, 2020, pp. 9213–9222.

[20] Y. Wang, R. Liu, D. Lin, D. Chen, P. Li, Q. Hu, and C. Chen, "Coarse-to-fine: Progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis." *TNNLS*, vol. 34, no. 2, pp. 761–774, 2023.

[21] R. Feng, X. Zheng, T. Gao, J. Chen, W. Wang, D. Chen, and J. Wu, "Interactive few-shot learning: Limited supervision, better medical image segmentation," *TMI*, vol. 40, no. 10, pp. 2575–2588, 2021.

[22] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICLR*, 2017, pp. 1126–1135.

[23] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*, 2019, pp. 10 657–10 665.

[24] D. Kang, H. Kwon, J. Min, and M. Cho, "Relational embedding for few-shot classification," in *ICCV*, 2021, pp. 8822–8833.

[25] C. Chen, K. Li, W. Wei, J. Zhou, and Z. Zeng, "Hierarchical graph neural networks for few-shot learning," *TCSVT*, 2021.

[26] M. Zhang, S. Huang, W. Li, and D. Wang, "Tree structure-aware few-shot image classification via hierarchical aggregation," in *ECCV*, 2022, pp. 453–470.

[27] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *ICLR*, 2018.

[28] S. Lu, H. Ye, and D. Zhan, "Tailoring embedding function to heterogeneous few-shot tasks by global and local feature adaptors," in *AAAI*, 2021, pp. 8776–8783.

[29] B. Oreshkin, P. Rodriguez, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *NeurIPS*, 2018, pp. 719–729.

[30] L. Bertinetto, J. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *ICLR*, 2018.

[31] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *CVPR*, 2019, pp. 1–10.

[32] Y. Tian, Y. Wang, D. Krishnan, J. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *ECCV*, 2020, pp. 266–282.
[33] S. Baik, J. Choi, H. Kim, D. Cho, J. Min, and K. Lee, "Meta-learning with task-adaptive loss function for few-shot learning," in *ICCV*, 2021, pp. 9465–9474.
[34] H. Huang, Z. Wu, W. Li, J. Huo, and Y. Gao, "Local descriptor-based multi-prototype network for few-shot learning," *PR*, vol. 116, p. 107935, 2021.
[35] 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 *AAAI*, vol. 35, no. 10, 2021, pp. 8635–8643.
[36] Y. Li, H. Li, H. Chen, and C. Chen, "Hierarchical representation based query-specific prototypical network for few-shot image classification," *arXiv preprint arXiv:2103.11384*, 2021.
[37] S. Gidaris and N. Komodakis, "Generating classification weights with gnn denoising autoencoders for few-shot learning," in *CVPR*, 2019, pp. 21–30.
[38] J. Yang, H. Yang, and L. Chen, "Towards cross-granularity few-shot learning: coarse-to-fine pseudo-labeling with visual-semantic meta-embedding," in *ACM MM*, 2021, pp. 3005–3014.
[39] Y. Su, H. Zhao, and Y. Lin, "Few-shot learning based on hierarchical classification via multi-granularity relation networks," *IJAR*, vol. 142, pp. 417–429, 2022.
[40] G. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classi?cation," in *ICLR*, 2020.
[41] W. Xu, Y. Xu, H. Wang, and Z. Tu, "Attentional constellation nets for few-shot learning," in *ICLR*, 2021.
[42] H. Chen, H. Li, Y. Li, and C. Chen, "Sparse spatial transformers for few-shot learning," *Science China Information Sciences*, 2023.
[43] A. Ravichandran, R. Bhotika, and S. Soatto, "Few-shot learning with embedded class models and shot-free meta training," in *ICCV*, 2019, pp. 331–339.
[44] W. Chen, Y. Liu, Z. Kira, Y. Wang, and J. Huang, "A closer look at few-shot classification," in *ICLR*, 2019.
[45] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, and V. Balasubramanian, "Charting the right manifold: Manifold mixup for few-shot learning," in *WACV*, 2020, pp. 2218–2227.
[46] Z. Xue, L. Duan, W. Li, L. Chen, and J. Luo, "Region comparison network for interpretable few-shot image classification," *arXiv preprint arXiv:2009.03558*, 2020.
[47] L. Qiao, Y. Shi, J. Li, Y. Wang, T. Huang, and Y. Tian, "Transductive episodic-wise adaptive metric for few-shot learning," in *ICCV*, 2019, pp. 3603–3612.

**Yuling Su** is currently a M.E. candidate with the School of Computer Science, Minnan Normal University, Fujian, China. Her research interests include the few-shot learning and hierarchical classification application in machine learning and granular computing.

**Hong Zhao** received the Ph.D degree from Tianjin University, Tianjin, China, in 2019. She received her M.S. degree from Liaoning Normal University, Dalian, China, in 2006. She is currently a Professor of the School of Computer Science and the Key Laboratory of Data Science and Intelligence Application, Minnan Normal University, Zhangzhou, China. She has authored over 50 journal and conference papers in the areas of granular computing based machine learning and cost-sensitive learning. Her current research interests include rough sets, granular computing, and data mining for hierarchical classification.

**Yifeng Zheng** received Ph.D degree in computer technology from China University of Petroleum-Beijing, Bejing, China in 2020. In 2004, he joined the faculty of School of Computer Science, Minnan Normal University, Zhangzhou, China. His research interests include artificial intelligence, machine learning and network communications.

**Yu Wang** received the B.S. degree in communication engineering, the M.S. degree in software engineering, and Ph.D. degree in computer applications and techniques from Tianjin University in 2013 and 2016, and 2020, respectively. He is currently an assistant professor of Tianjin University and was an outstanding visitor scholar of the University of Waterloo in 2019. His research interests focus on hierarchical learning and large-scale classification in industrial scenarios and computer vision applications, data mining, and machine learning. He has published many peer-reviewed papers in world-class conferences and journals, such as IEEE TFS, TNNLS, TCYB, TKDE, etc.