Few-shot Classification via Ensemble Learning with Multi-Order Statistics

Sai Yang[1] , **Fan Liu**[2,3*] , **Delong Chen**[2] and **Jun Zhou**[4]
[1]School of Electrical Engineering, Nantong University, Nantong, China
[2]College of Computer and Information, Hohai University, Nanjing, China
[3]Science and Technology on Underwater Vehicle Technology Laboratory, Harbin Engineering University, Harbin, China
[4]School of Information and Communication Technology, Griffith University, Queensland, Australia
fanliu@hhu.edu.cn

Abstract

Transfer learning has been widely adopted for few-shot classification. Recent studies reveal that obtaining good generalization representation of images on novel classes is the key to improving the few-shot classification accuracy. To address this need, we prove theoretically that leveraging ensemble learning on the base classes can correspondingly reduce the true error in the novel classes. Following this principle, a novel method named Ensemble Learning with Multi-Order Statistics (ELMOS) is proposed in this paper. In this method, after the backbone network, we use multiple branches to create the individual learners in the ensemble learning, with the goal to reduce the storage cost. We then introduce different order statistics pooling in each branch to increase the diversity of the individual learners. The learners are optimized with supervised losses during the pre-training phase. After pre-training, features from different branches are concatenated for classifier evaluation. Extensive experiments demonstrate that each branch can complement the others and our method can produce a state-of-the-art performance on multiple few-shot classification benchmark datasets.

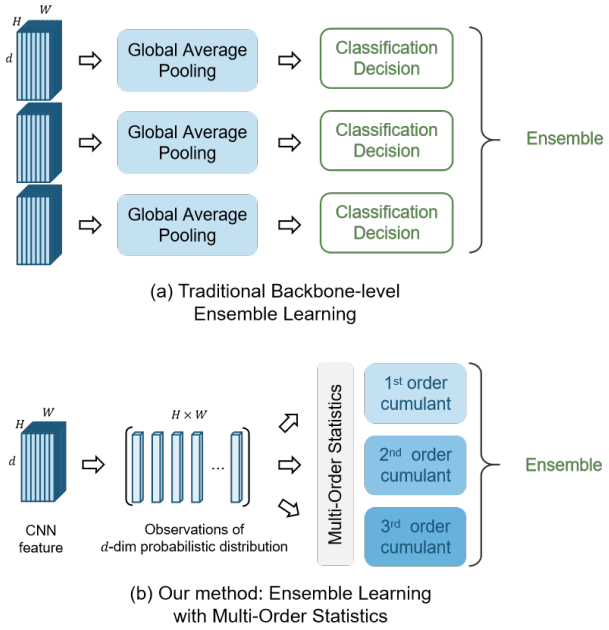


Figure 1: (a) The traditional methods often use different backbone networks as individuals, which significantly increases the computation and storage costs. (b) Our method takes the same backbone and equips different branches with multi-order statistics as learning individuals. They are parameter-free and trained jointly, and do not require extra model size and computation time.

1 Introduction

Few-shot Classification (FSC) is a promising direction in alleviating the labeling cost and bridging the gap between human intelligence and machine models. It aims to accurately differentiate novel classes with only a few labeled training samples. Due to limited supervision from novel classes, an extra base set with abundant labeled samples is often used to improve the classification performance. According to the adopted training paradigms, FSC methods can be roughly divided into meta-learning-based [Finn *et al.*, 2017; Snell *et al.*, 2017] and transfer-learning-based [Chen *et al.*, 2019; Liu *et al.*, 2020; Afrasiyabi *et al.*, 2020]. The first type takes the form of episodic training, in which subsets of data are sampled from the base set to imitate the meta-test setting. Since sampling does not cover all combinations, this paradigm cannot fully utilize the information provided by the base set. In contrast, the transfer-learning takes the base set as a whole, so it avoids the drawback of meta-learning and achieves better performance. Many effective regularization techniques have been exploited in transfer-learning, for example, manifold mixup [Mangla *et al.*, 2020], self-distillation [Tian *et al.*, 2020], and self-supervised learning [Zhang *et al.*, 2020b], which leads to significant improvement on the generalization of image representations and the FSC performance.

Ensemble learning combines multiple learners to solve the same problem and exhibits better generalization performance than any individual learners. When combining ensemble learning with deep Convolutional Neural Networks (CNN), the new paradigm usually requires large-scale training data for classification tasks [Horváth *et al.*, 2021; Agarwal *et al.*, 2021], making it challenging to be adopted for

FSC. Recently, two notable studies [Dvornik *et al.*, 2019; Bendou *et al.*, 2022] employed an ensemble of deep neural networks for FSC tasks under either a meta-learning or a transfer-learning setting. They demonstrated that ensemble learning is also applicable to FSC. Yet, these works are still preliminary and lack a theoretical analysis to explain the underlying reason behind the promising performance. To address this challenge, we provide an FSC ensemble learning theorem for the transfer-learning regime. Its core idea is a tighter expected error bound on the novel classes, in which the expected error on the novel classes can be reduced by implementing ensemble learning on the base classes, given the base classes-novel classes domain divergence.

The generalization ability of ensemble learning is strongly dependent on generating diverse individuals [Yang *et al.*, 2013]. As shown in Figure 1 (a), traditional methods often use different backbone networks as individuals, which significantly increases the computation and storage costs. Our work finds that different-order statistics of the CNN features are complementary to each other, and integrating them can better model the whole feature distribution. Based on this observation, we develop a parameter-free ensemble method, which takes the same backbone and equips different branches with multi-order statistics as learning individuals. We name this method Ensemble Learning with Multi-Order Statistics (ELMOS), as shown in Figure 1 (b). The main contributions of this paper are summarized as follows:

- To our knowledge, this is the first theoretical analysis to guide ensemble learning in FSC. The derived theorem proves a tighter expected error bound is available on novel classes.
- We propose an ensemble learning method by adding multiple branches at the end of the backbone networks, which can significantly reduce the computation time of the training stage for FSC.
- This is the first time that multi-order statistics is introduced to generate different individuals in ensemble learning.
- We conduct extensive experiments to validate the effectiveness of our method on multiple FSC benchmarks.

2 Related Work

2.1 Few-shot Classification

According to how the base set is used, FSC methods can be roughly categorized into two groups, meta-learning-based [Bertinetto *et al.*, 2019; Zhang *et al.*, 2020a] and transfer-learning-based [Chen *et al.*, 2019; Liu *et al.*, 2020]. Meta-learning creates a set of episodes to simulate the real FSC test scenarios and simultaneously accumulate meta-knowledge for fast adaptation. Typical meta-knowledge includes optimization factors such as initialization parameters [Finn *et al.*, 2017] and task-agnostic comparing ingredients of feature embedding and metric [Snell *et al.*, 2017; Wertheimer *et al.*, 2021]. Recent literature on transfer learning [Tian *et al.*, 2020; Chen *et al.*, 2019] questioned the efficiency of the episodic training in meta-learning, and alternatively used all base samples to learn an off-the-shelf feature extractor and rebuilt a classifier for novel classes. Feature representations play an important role in this regime [Tian *et al.*, 2020]. To this end, regularization techniques such as negative-margin softmax loss and manifold mixup [Liu *et al.*, 2020; Mangla *et al.*, 2020] have been adopted to enhance the generalization ability of cross-entropy loss. Moreover, self-supervised [Zhang *et al.*, 2020b; Rajasegaran *et al.*, 2020] and self-distillation [Ma *et al.*, 2019; Zhou *et al.*, 2021] methods have also shown promising performance in transfer-learning. To this end, supervised learning tasks can be assisted by several self-supervised proxy tasks such as rotation prediction and instance discrimination [Zhang *et al.*, 2020b], or by adding an auxiliary task of generating features during the pre-training [Xu *et al.*, 2021b]. When knowledge distillation is adopted, a high-quality backbone network can be evolved through multiple generations by a born-again strategy [Rajasegaran *et al.*, 2020]. All these methods suggest the importance of obtaining generalization representations, and we will leverage ensemble learning to achieve this goal.

2.2 Ensemble Learning

Ensemble learning builds several different individual learners based on the same training data and then combines them to improve the generalization ability of the learning system over any single learner. This learning scheme has shown promising performance on traditional classification tasks with deep learning on large-scale labeled datasets [Horváth *et al.*, 2021; Agarwal *et al.*, 2021]. Recently, ensemble learning for FSC methods has been presented. For example, [Dvornik *et al.*, 2019] combined an ensemble of prototypical networks through deep mutual learning under a meta-learning setting. [Bendou *et al.*, 2022] reduced the capacity of each backbone in the ensemble and pre-trained them one by one with the same routine. However, the size of the ensemble learner increased for inference in the former work, while the latter required extra time to pre-train many learning individuals. Therefore, it still lacks efficient designs for learning individuals in FSC ensemble learning. Moreover, these works did not involve any theoretical analysis of the underlying mechanism of ensemble learning in FSC. In this paper, we investigate why ensemble learning works well in FSC under the transfer-learning setting. Based on the analysis, we propose an efficient learning method using a shared backbone network with multiple branches to generate learning individuals.

2.3 Pooling

Convolutional neural network models progressively learn high-level features through multiple convolution layers. A pooling layer is often added at the end of the network to output the final feature representation. To this end, Global Average Pooling (GAP) is the most popular option, however, it cannot fully exploit the merits of convolutional features because it only calculates the 1^{st}-order feature statistics. Global Covariance Pooling (GCP) such as DeepO^{2}P [Ionescu *et al.*, 2015] explores the 2^{nd}-order statistic by normalizing the covariance matrix of the convolutional features, which has achieved impressive performance gains over the classical GAP in various computer vision tasks. Further research

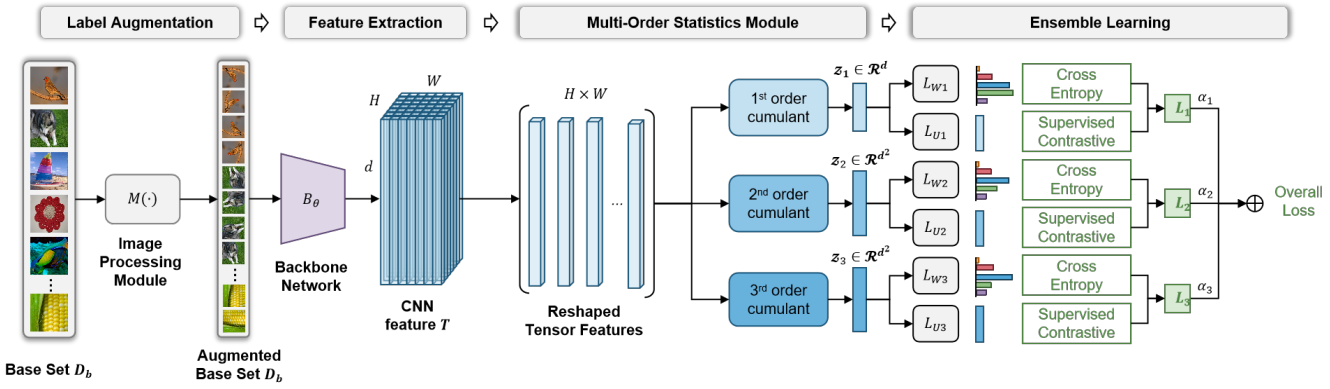


Figure 2: An overview of our framework. The images from S_b are augmented by the image processing module and fed into the backbone for feature extraction. The CNN features from the backbone are then reshaped into the matrix, which is used to calculate multi-order statistics to equip different branches. Ensemble learning is implemented by the linear combination of multiple branches during the pre-training phase.

shows that using richer statistics may lead to further possible improvement. For example, Kernel Pooling [Cui *et al.*, 2017] generates high-order feature representations in a compact form. However, a certain order statistic can only describe partial characteristics of the feature vector from the view of the characteristic function of random variables. For example, the first- and second-order statistics can completely represent their statistical characteristic only for the Gaussian distribution. Therefore, higher-order statistics are still needed for the non-Gaussian distributions, which are more ubiquitous in many real-world applications. This motivates us to calculate multi-order statistics to retain more information on features.

3 The Proposed Method

Here we present the proposed method. We start with a formal definition of FSC, and then present a theorem on FSC ensemble learning. This theorem leads to the development of an ensemble learning approach with multi-order statistics.

3.1 Theory Foundation

Under the standard setting of few-shot classification, three sets of data with disjoint labels are available, i.e., the base set S_b, the validation set S_{val} and the novel set S_n. In the context of transfer-learning, S_b is used for pre-training a model to well classify the novel classes in S_n, with the hyper-parameters tuned on S_{val}. Let $S_b = \{(x_i, y_i)\}_{i=1}^{N_b}$ denotes the source domain with N_b labelled samples and S_n denotes the target domain labelled with K samples in each episode, where $N_b >> K$. Let the label function of S_b and S_n be f_b and f_n, respectively. During the pre-training, a learner h is obtained to approximate the optimal mapping function h^* based on all N_b training samples in S_b from all possible hypotheses $\mathcal{H}$. When ensemble learning is introduced into the pre-training, several learners denoted as $\{h_o\}_{o=1}^{O}$ can be obtained. With the ensemble technique of weighted averaging, the final learner $\overline{h}$ is produced as:

$$\overline{h} = \sum_{o=1}^{O} \alpha_o h_o, \tag{1}$$

where α_o is the weight parameter. There is a domain shift between the base and novel classes [Tseng *et al.*, 2020], and we use the L_1 distance [Kifer *et al.*, 2004] to measure the domain divergence between S_b and S_n:

$$\mathcal{D}(S_b, S_n) = \int |\eta_b(x) - \eta_n(x)| \left|\overline{h}(x) - f_n(x)\right| dx, \tag{2}$$

where $\eta_b(x)$ and $\eta_n(x)$ is the density functions of S_b and S_n respectively.

Theorem 1 (FSC Ensemble Learning) *Let $\mathcal{H}$ be a hypothesis space, for any $h \in \{h_o\}_{o=1}^{O} \in \mathcal{H}$ is learned from S_b, and $\overline{h} = \sum_{o=1}^{O} \alpha_o h_o \in \mathcal{H}$, the expected error on S_n respectively with $\overline{h}$ and h holds the following relationship:*

$$\begin{aligned} e_n(\overline{h}) &\leq e_b(\overline{h}) + \underbrace{\mathcal{D}(S_b, S_n)}_{(S_b\text{-}S_n)\ divergence} + \lambda \\ &\leq e_b(h) + \underbrace{\mathcal{D}(S_b, S_n)}_{(S_b\text{-}S_n)\ divergence} + \lambda, \end{aligned}$$

where $\lambda = E_{X \in S_b} |f_n(x) - f_b(x)|$ is a constant, $e_n(\overline{h})$ is the expected error on S_n with $\overline{h}$, $e_b(h)$ is the expected error on S_b with h, $e_b(\overline{h})$ is the expected error on S_b with $\overline{h}$.

The proof is provided in the Supplementary Material.

Remark 1 *The core idea of Theorem 1 is to define a tighter expected error bound on the novel classes with the learned mapping function in the form of ensemble learning during the pre-training. Theorem 1 tells that the true error on the novel classes can be reduced by implementing ensemble learning on the base classes, given the domain divergence between the novel class and base class. This can well explain the effectiveness of ensemble learning in few-shot classification, in which multiple learners are assembled to enhance the generalization on the base set, resulting in better performance in novel classes.*

3.2 FSC via Ensemble Learning with Multi-order Statistics

Overview

Our method employs the transfer-learning paradigm in a two-phase manner. In the first phase, a good feature extractor is

pre-trained on the base set. In the second phase, FSC evaluation is done on the novel set with the pre-trained feature extractor. Following Theorem 1, we introduce ensemble learning in the first phase to improve the FSC performance. The key to this phase is to effectively train multiple diverse individuals. Different from the previous works [Dvornik *et al.*, 2019; Bendou *et al.*, 2022] that use many different networks as individuals, we add multiple branches after the backbone network to create individuals for reducing training costs. Each branch calculates different-order statistics for pooling to highlight the discrepancy between the individuals. This step is optimized by supervised losses. After pre-training, features from different branches are concatenated for FSC evaluation. We name this method as Ensemble Learning with multi-Order Statistics (ELMOS) for FSC. An overview of ELMOS is shown in Figure 2, and a flow description of ELMOS is given in Algorithm 1.

Pre-training via Multi-order Statistics

The proposed model architecture mainly consists of the following four components: an image processing module, the backbone network, a multi-order statistics module, and a supervised classifier module. The image processing module is denoted as $M(\cdot)$, which performs transformation of multi-scale rotation to augment the original base set and their label space. The backbone network is denoted as $B_\theta(\cdot)$ and parameterized by θ, which converts each image into a tensor of size $H \times W \times d$. The multi-order statistics module module is denoted as $S(\cdot)$, which maps the tensor from the backbone into multiple feature representations to generate individual learners for ensemble learning. The supervised classifier module is composed of softmax classifiers $L_W(\cdot)$ and the projectors $L_U(\cdot)$ with parameter matrices W and U, respectively, which are used to build the supervised losses for pre-training.

Given L samples be randomly sampled from S_b with C_b classes, in which an image and its corresponding label are denoted as (x_i, y_i), $y_i \in \{1, 2, ...C_b\}$. $M(\cdot)$ scales the images with the aspect-ratio of 2:3 and rotates the images with $\{0°, 90°, 180°, 270°\}$ under both the new and the original scales, resulting in eight times expansion of training samples. Feed x_i into B_θ to produce a tensor feature of $T_i = B_\theta(x_i) \in \mathscr{R}^{H\times W\times d}$. Next, we reshape the tensor T_i into the matrix $T_i \in \mathscr{R}^{HW\times d}$, and view each row vector in the matrix $t_j \in \mathscr{R}^d$ as an observation of the random variable of $t \in \mathscr{R}^d$. When $d = 1$, the first characteristic function of variable t in the Laplace operator is given by:

$$\phi(s) = \int_{-\infty}^{+\infty} f(t)e^{st}dt = \int_{-\infty}^{+\infty} e^{st}dF(t), \tag{3}$$

where $f(t)$ and $F(t)$ are the density function and distribution function of t, respectively. Let $\psi(s) = ln\phi(s)$ be the second characteristic function of the random variable t.

Theorem 2 (The Inversion Formula for Distributions) *Let t be a random variable with distribution function $F(t)$ and characteristic function $\phi(s)$. For $a, b \in C(F)$ and $a < b$,*

$$F(b) - F(a) = \lim_{c\to\infty} \frac{1}{2\pi} \int_{-c}^{c} \frac{e^{-sa} - e^{-sb}}{s} \phi(s)ds.$$

Corollary 1 (Uniqueness) *If two distributions of $F_1(t)$ and $F_2(t)$ are identical, then the corresponding characteristic functions $\psi_1(s)$ and $\psi_2(s)$ are identical.*

See proof of Theorem 2 and Corollary 1 in [Shiryaev, 2016]. From Theorem 2 and Corollary 1, we can see that there is a one-to-one correspondence between the characteristic function and the probability density function such that the characteristic function can completely describe a random variable.

The o^{th}-order cumulant of the random variable t is defined as the o^{th} derivative of function $\psi(s)$ at the origin, which is:

$$c_o = \left.\frac{d^o\psi(s)}{ds^o}\right|_{s=0}. \tag{4}$$

Then the Taylor series expansion of function $\psi(s)$ at the origin with respect to s yields:

$$\psi(s) = c_1 s + \frac{1}{2}c_2 s^2 + ... + \frac{1}{o!}c_p s^o + R_s(s^o), \tag{5}$$

where $R_s(s^o)$ is the remainder term. It can be seen from Equation (5) that the o^{th}-order cumulant of t is the coefficient of the term s^o in Equation (5).

Proposition 1 *Consider a Gaussion distribution $f(t)$ with mean μ and variance Σ^2 for the random variable t, its second characteristic function is:*

$$\psi(s) = \mu s + \frac{1}{2}\Sigma^2 s^2.$$

Consequently, the cumulant of the random variable t are:

$$c_1 = \mu, c_2 = \Sigma^2, c_o = 0 \quad (o = 3, 4, ...).$$

Remark 2 *Proposition 1 implies that for Gaussian signals only, the cumulants are identically zero when the order is greater than 2. Please note this conclusion can be naturally extended to the scenario of multivariate variables when $d > 1$. For the random variables with Gaussian distribution, the first and second-order statistics can completely represent their statistical characteristics. However, the non-Gaussian signals are more common in real-world applications. In this case, higher-order statistics also contain a lot of useful information. Therefore, we propose a multi-order statistics module consisting of multiple branches, each equipped with different order statistics of the tensor feature T_i.*

In particular, we employ three branches in the multi-order statistics module, which respectively calculate three orders cumulants of the variable t with the observations in T_i. The specific formulation of the 1^{st}-order, 2^{nd}-order and 3^{rd}-order cumulants of t are expressed as:

$$\begin{aligned}
c_{i1} &= \frac{1}{H\times W}\sum_{j=1}^{H\times W} t_j \quad c_{i1} \in \mathscr{R}^d, \\
c_{i2} &= \frac{1}{H\times W}\sum_{j=1}^{H\times W} (t_j - c_{i1})(t_j - c_{i1})^T \quad c_{i2} \in \mathscr{R}^{d\times d}, \\
c_{i3} &= \frac{1}{H\times W}\sum_{j=1}^{H\times W} \frac{(t_j - c_{i1})^2(t_j - c_{i1})^T}{c_{i2}^2 c_{i2}^T} \quad c_{i3} \in \mathscr{R}^{d\times d}.
\end{aligned} \tag{6}$$

As c_{i2} and c_{i3} are $d \times d$ matrices, we flatten them into d^2-dimensional vectors and finally get the feature representations of z_{i1}, z_{i2} and z_{i3}. We use these three features as individuals in ensemble learning, which respectively pass through their corresponding softmax classifier $L_W(\cdot)$ and projectors $L_U(\cdot)$. So the o-th ($o = 1, 2, 3$) outputs are:

$$P_{ij}^{o} = L_{Wo}(z_{io}) = \frac{exp({z_{i0}}^T w_{oj})}{\sum_{j=1}^{8C_b} exp({z_{i0}}^T w_{oj})}, \quad u_{io} = \|L_{Uo}(z_{io})\| = \left\|{z_{io}}^T U_o\right\|, \tag{7}$$

where $L_{Wo}(\cdot)$ is the o-th softmax classifier with the parameter matrix of W_o, w_{oj} is the j-th component of W_o. $L_{Uo}(\cdot)$ is the o-th projector with the parameter matrix U_o. P_{ij}^{o} is the j-th component of the output probability from the o-th softmax classifier. u_{io} is the output vector from the o-th projector. We simultaneously employ Classification-Based (CB) loss of cross-entropy and Similarity-Based (SB) loss of supervised contrastive in supervised learning for each individual [Scott *et al.*, 2021]. These two losses are formulated as:

$$L_{CB}^{o}(\theta, W_o) = -\sum_{i=1}^{8L}\sum_{j=1}^{8C_b} y_{ij} log P_{ij}^{o},$$

$$L_{SB}^{o}(\theta, U_o) = -\sum_{i=1}^{8L} log \sum_{q \in Q(u_{io})} \frac{exp(u_{io} \cdot u_{qo}/\tau)}{\sum_{a=1}^{8L} exp(u_{ao} \cdot u_{qo}/\tau)}, \tag{8}$$

where y_{ij} is the j-th component of label y_i, τ is a scalar temperature parameter. $Q(u_{io})$ is the positive sample set, in which each sample has the same label as u_{io}. u_{qo} is the q-th sample in $Q(u_{io})$. Then the learning objective function for the o-th individual is:

$$L_o(\theta, W_o, U_o) = L_{CB}^{o}(\theta, W_o) + L_{SB}^{o}(\theta, U_o). \tag{9}$$

The overall loss function with ensemble learning is:

$$L_{overall} = \sum_{o=1}^{O} \alpha_o L_o(\theta, W_o, U_o), \tag{10}$$

where α_o is a weight controlling the contribution of each individual in the ensemble learning. The pre-training adopts the gradient descent method to optimize the above loss function.

Few-shot Evaluation

The phase of few-shot evaluation still needs to construct a set of N-way K-shot FSC tasks, with a support set and a query set in each task. The support set randomly selects K samples from each of the N classes that are sampled from S_n, which is denoted as $S_p = \{x_s, y_s\}_{s=1}^{NK}$, where (x_s, y_s) is the s-th images and its corresponding label. The query set consists of the remaining images in these N classes, which is denoted as $S_q = \{x_q\}_{q=1}^{Q}$ with any image of x_q. After pre-training, we get rid of the softmax classifier $L_W(\cdot)$ and projectors $L_U(\cdot)$ and fix the backbone network $B_\theta(\cdot)$ and the multi-order statistics module module $S(\cdot)$. The support set S_p is input into $B_\theta(\cdot)$ and $S(\cdot)$ to produce the output features:

$$z_{so} = B_\theta \circ S(x_s) \quad (o = 1, 2, 3), \tag{11}$$

Algorithm 1: Ensemble Learning with multi-Order Statistics (ELMOS) for FSC

Input: Base set S_b, support set S_p, query set S_q; augmentation module $M(\cdot)$, backbone network $B_\theta(\cdot)$, multi-order statistics module $S(\cdot)$, softmax classifier L_{Wo}, projector L_{Uo} and logistic regression $g_\xi(\cdot)$; temperature parameter τ, weight α_o ($o = 1, 2, 3$).
Output: Final prediction of the query samples
Stage 1: Pre-training with ensemble learning
for *numbers of training epochs* **do**
  Sample a mini-batch with any image of $\{x_i, y_i\}$;
  Feed x_i into $T(\cdot)$ and $B_\theta(\cdot)$ to obtain feature map $T_i \in \mathscr{R}^{H \times W \times d}$;
  Pass T_i through $S(\cdot)$ to output features z_{io}, ($o = 1, 2, 3$);
  Pass z_{io} through L_{Wo} and L_{Uo} to get the output probability and projection feature ;
  Calculate optimization loss for each individual via Equation (9);
  Calculate overall loss for pre-training via Equation (10);
  Update the parameters of θ, W_o, U_o using SGD;
end
Stage 2: Few-shot evaluation
for *all iteration = 1, 2, ..., MaxIteration* **do**
  Feed $x_s \in S_p$ into $B_\theta(\cdot)$ and $S(\cdot)$ to output feature z_{so}, ($o = 1, 2, 3$);
  Concatenate z_{so} into the feature z_s to train the classifier of $g_\xi(\cdot)$;
end
Classify the query samples according to Equation (13).

where $\circ$ is the stack operator. The features z_{s1}, z_{s2}, z_{s3} are concatenated into a final expression of x_s:

$$z_s = con(z_{s1}, z_{s2}, z_{s3}), \tag{12}$$

where $con(\cdot)$ is the concatenated operator. A logistic regression classifier $g_\xi(\cdot)$ parameterized by ξ is then trained with z_s and its corresponding label y_s. The query image x_q is finally classified as:

$$\hat{y}_q = g_\xi(z_q), \tag{13}$$

where $\hat{y}_q$ is the inference label value of x_q.

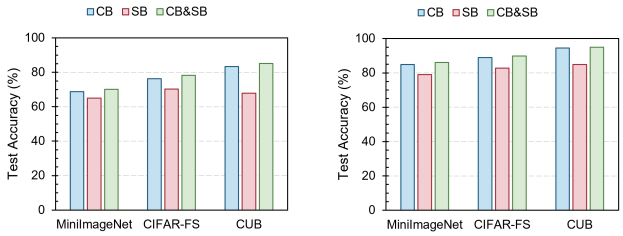
4 Experiments

4.1 Datasets

miniImageNet contains 600 images over 100 classes, which are divided into 64, 16 and 20 respectively for base, validation and novel sets. **tiredImageNet** consists of 779, 165 images belonging to 608 classes, which are further grouped into 34 higher-level categories with 10 to 30 classes per category. These categories are partitioned into 20 categories (351 classes), 6 categories (97 classes) and 8 categories (160 classes) respectively for base, validation and novel sets. **CIFAR-FS** is derived from CIFAR100 and consists of 100 classes with 600 images per class. The total

Method	Backbone	miniImageNet		CIFAR-FS		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
B_1	ResNet12	69.06±0.44	83.61±0.29	77.09±0.46	88.46±0.34	81.46±0.39	92.55±0.18
B_2	ResNet12	66.42±0.42	85.76±0.26	71.53±0.48	88.83 ±0.27	77.79±0.39	94.44±0.17
B_3	ResNet12	67.68±0.43	82.81±0.29	72.83±0.46	86.34±0.34	83.89±0.38	91.20±0.17
ELMOS	ResNet12	70.30±0.45	86.17±0.26	78.18±0.41	89.87±0.31	85.21±0.38	95.02±0.16

Table 1: Test accuracy (%) of each branch and their ensemble under 5-way 1-shot and 5-shot tasks on three datasets.



(a) 5-way 1-shot (b) 5-way 5-shot

Figure 3: Test accuracy (%) of the classification-based (CB) loss, similarity-based (SB) loss and their combination (CB&SB) under 5-way 1-shot and 5-way 5-shot tasks on three datasets.

Method	CUB	
	1-shot	5-shot
Meta-learning		
DeepEMD [Zhang *et al.*, 2020a]	75.65±0.83	88.69±0.50
BML [Zhou *et al.*, 2021]	76.21±0.63	90.45±0.36
RENet [Kang *et al.*, 2021]	79.49±0.44	91.11± 0.24
FPN[Wertheimer *et al.*, 2021]	83.55±0.19	92.92±0.10
IEPT [Zhang *et al.*, 2020b]	69.97±0.49	84.33±0.33
APP2S [Ma *et al.*, 2022b]	77.64±0.19	90.43±0.18
MFS [Afrasiyabi *et al.*, 2022]	79.60±0.80	90.48±0.44
DeepBDC [Xie *et al.*, 2022]	84.01±0.42	94.02±0.24
HGNN [Yu *et al.*, 2022]	78.58±0.20	90.02±0.12
INSTA[Ma *et al.*, 2022a]	75.26±0.31	88.12±0.54
Transfer-learning		
Neg-Cosine [Liu *et al.*, 2020]	72.66±0.85	89.40±0.43
S2M2 [Mangla *et al.*, 2020]	80.68±0.81	90.85±0.44
DC-LR[Yang *et al.*, 2021]	79.56±0.87	90.67±0.35
CCF [Xu *et al.*, 2021b]	81.85±0.42	91.58±0.32
ELMOS (ours)	85.21±0.38	95.02±0.16

Table 2: Comparison of results against state-of-the-art methods on CUB dataset.The top three results are marked in red, blue and green.

classes are split into 64, 16 and 20 for base, validation and novel sets. **Caltech-UCSD Bird-200-2011(CUB)** has a total number of 11,788 images over 200 bird species. These species are divided into 100, 50, and 50 for the base, validation and novel sets, respectively.

4.2 Implementation Details

In the experiments, we primarily used ResNet12 architecture with 4 residual blocks. Each block had 3 convolutional layers with 3×3 kernels. The number of kernels for the 4 blocks was 64, 160, 320, and 640, respectively. A max-pooling layer was added at the end of the first three blocks. The last block was branched with three pooling layers, which respectively modeled different statistical representations of the images. We opted for the SGD optimizer with a momentum of 0.9 and a weight decay of 5e-4. The learning rate was initialized to be 0.025. We trained the network for 130 epochs with a batch size of 32 in all the experiments. For miniImageNet, tiredImageNet and CIFAR-FS, the learning rate was reduced by a factor of 0.2 at the 70-th and 100-th epoch. For CUB, the learning rate was reduced by a factor of 0.2 for every 15 epochs after the 75-th epoch. We randomly sampled 2,000 episodes from S_n with 15 query samples per class for both 5-way 1-shot and 5-shot evaluations, to produce the mean classification accuracy as well as the 95% confidence interval.

4.3 Ablation Studies

The effectiveness of our method is attributed to the ensemble of different branches equipped with multi-order statistics. In this section, we conducted ablation studies to analyze the effect of the 1^{st}-order, 2^{nd}-order and, 3^{rd}-order statistical pooling and their combination on the miniImageNet, CIFAR-FS and CUB datasets. Above methods are respectively denoted as B_1, B_2, B_3,and ELMOS. Their accuracies under 5-way 1-shot and 5-shot tasks on three datasets are shown in Table 1. From the results, we can see that: (1) On all three datasets, the test accuracy of B_1 and B_3 is higher than B_2 under the 1-shot task, but the test accuracy of B_2 is higher than B_1 and B_3 under the 5-shot task. The above phenomenon shows that different order statistics provide different information about the images. (2) The test accuracy of ELMOS is higher than B_1, B_2 and B_3 under both 1-shot and 5-shot tasks, which illustrates that different order statistics complement each other. Combing them can bring more useful information for classification, resulting in higher classification performance.

For each individual in the ensemble learning, the optimization is cooperatively accomplished by the Classification-Based (CB) loss and Similarity-Based (SB) loss [Scott *et al.*, 2021]. Hence, we conducted ablation experiments to analyze the contribution of each loss on three benchmark datasets: miniImageNet, CIFAR-FS and CUB. Subsequently, we pretrained the model respectively with CB and SB loss alone and their combination, resulting in three methods denoted as CB, SB and CB&SB. The test accuracies under different methods are shown in Figure 3. The test results show that the accuracy of CB&SB is higher than CB and SB, which implies that both CB and SB losses play important roles in our method.

4.4 Comparison with the Most Related Method

Our method is most related to EASY [Bendou *et al.*, 2022], which is also a FSC ensemble learning method in context of

Method	Backbone	Venue	miniImageNet		tiredImageNet		CIFAR-FS	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Meta-learning								
DeepEMD[Zhang *et al.*, 2020a]	ResNet12	CVPR'20	65.91±0.82	82.41± 0.56	71.16±0.87	86.03±0.58	-	-
CC+rot [Gidaris *et al.*, 2019]	ResNet12	CVPR'20	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36	76.09±0.30	87.83±0.21
BML [Zhou *et al.*, 2021]	ResNet12	ICCV'21	67.04±0.63	83.63±0.29	68.99±0.50	85.49±0.34	73.45±0.47	88.04±0.33
RENet [Kang *et al.*, 2021]	ResNet12	ICCV'21	67.60±0.44	82.58±0.30	71.61±0.51	85.28±0.35	74.51±0.46	86.60±0.32
MeTAL[Baik *et al.*, 2021]	ResNet12	CVPR'21	66.61±0.28	81.43±0.25	70.29±0.40	86.17±0.35	-	-
IEPT [Zhang *et al.*, 2020b]	ResNet12	ICLR'21	67.05±0.44	82.90±0.30	72.24±0.50	86.73±0.34	-	-
DAN [Xu *et al.*, 2021a]	ResNet12	CVPR'21	67.76±0.46	82.71±0.31	71.89±0.52	85.96±0.35	-	-
APP2S [Ma *et al.*, 2022b]	ResNet12	AAAI'22	66.25±0.20	83.42±0.15	72.00±0.22	86.23±0.15	73.12 ±0.22	85.69±0.16
DeepBDC [Xie *et al.*, 2022]	ResNet12	CVPR'22	67.34±0.43	84.46±0.28	72.34±0.49	87.31±0.32	-	-
MFS [Afrasiyabi *et al.*, 2022]	ResNet12	CVPR'22	68.32±0.62	82.71±0.46	73.63±0.88	87.59±0.57	-	-
TPMN[Wu *et al.*, 2021]	ResNet12	CVPR'22	67.64±0.63	83.44±0.43	72.24± 0.70	86.55 ± 0.63	-	-
HGNN [Yu *et al.*, 2022]	ResNet12	AAAI'22	67.02±0.20	83.00±0.13	72.05±0.23	86.49±0.15	-	-
MTR[Bouniot *et al.*, 2022]	ResNet12	ECCV'22	62.69± 0.20	80.95±0.14	68.44 ±0.23	84.20 ±0.16	-	-
Transfer-learning								
Neg-Cosine [Liu *et al.*, 2020]	WRN28	ECCV'20	61.72±0.81	81.79±0.55	-	–	-	
RFS [Tian *et al.*, 2020]	WRN28	ECCV'20	64.82±0.60	82.14±0.43	71.52±0.69	86.03±0.49	-	-
CBM [Wang *et al.*, 2020]	ResNet12	MM'20	64.77±0.46	80.50±0.33	71.27±0.50	85.81±0.34	-	-
SKD [Rajasegaran *et al.*, 2020]	ResNet12	Arxiv'21	67.04±0.85	83.54±0.54	72.03±0.91	86.50±0.58	76.9±0.9	88.9±0.6
IE [Sung *et al.*, 2021]	ResNet12	CVPR'21	67.28±0.80	84.78±0.33	72.21±0.90	87.08±0.58	77.87±0.85	89.74±0.57
PAL [Ma *et al.*, 2019]	ResNet12	ICCV'21	69.37±0.64	84.40±0.44	72.25±0.72	86.95±0.47	77.1±0.7	88.0±0.5
CCF[Xu *et al.*, 2021b]	ResNet12	CVPR'22	68.88±0.43	84.59±0.30	-	-	-	-
ELMOS (ours)	ResNet12	-	70.30±0.45	86.17±0.26	73.84±0.49	87.98±0.31	78.18±0.41	89.87±0.31

Table 3: Comparison of results against state-of-the-art methods on miniImageNet, tiredImageNet, and CIFAR-FS dataset. '-' means the results were not provided by the authors. The top three results are marked in red, blue and green, respectively.

Method	CIFAR-FS		CUB	
	1-shot	5-shot	1-shot	5-shot
EASY	75.24±0.20	88.38±0.14	77.97±0.20	91.59±0.10
ELMOS	78.18±0.41	89.87±0.31	85.21±0.38	95.02±0.16

Table 4: Comparison of results with the most related method under 5-way 1-shot and 5-shot tasks on CIFAR-FS and CUB.

transfer learning. The comparison of results between them on CIFAR-FS and CUB datasets is shown in Table 4. From the results, we can see that our method beats EASY by a very large margin under both 1-shot and 5-shot tasks. Please note that our method is more efficient that EASY, because EASY needs to pre-train multiple individual networks, which spends much more pre-training time than our method.

4.5 Comparison with State-of-the-Art Methods

We compare the performance of our method with several state-of-the-art methods. As shown in Table 2 and Table 3, we can see the performance of our method ranks at the top under both 1-shot and 5-shot tasks on CUB. Specifically, our method exceeds the second-best model DeepBDC by 1.2% and 1.0% respectively in 1-shot and 5-shot settings. From Table 3, we can see that our method beats state-of-the-art methods under both 5-way 1-shot and 5-way 5-shot tasks on the dataset of miniImageNet, tiredImagegNet, and CIFAR-FS. Specifically, on miniImageNet, PAL and IE behave the second best respectively in 1-shot and 5-shot settings. Our method beats them by 0.93% and 1.39%. On tiredImageNet, our method outperforms the second-best MFS by 0.21% and 0.39% respectively in 1-shot and 5-shot settings. On CIFAR-FS, our method achieves 0.31% and 0.13% improvement over IE for 1-shot and 5-shot respectively. In brief, our method consistently outperforms the state-of-the-art FSC methods under both 5-way 1-shot and 5-way 5-shot tasks on multiple datasets.

5 Conclusion

This paper analyzes the underlying work mechanism of ensemble learning in FSC. A theorem is provided to illustrate that the true error on the novel classes can be reduced with ensemble learning on the base set, given the domain divergence between the base and the novel classes. Multi-order statistics on image features are further introduced to produce learning individuals to get an effective ensemble learning design. Comprehensive experiments on multiple benchmarks have illustrated that different-order statistics can generate diverse learning individuals due to their complementarity.

Acknowledgments

This work was partially supported by Joint Fund of Ministry of Education for Equipment Pre-research(8091B022123), Research Fund from Science and Technology on Underwater Vehicle Technology Laboratory (2021JCJQ-SYSJJ-LB06905), Key Laboratory of Information System Requirements, No: LHZZ 2021-M04, Water Science and Technology Project of Jiangsu Province under grant No.2021063, Qinglan Project of Jiangsu Province.

References

[Afrasiyabi *et al.*, 2020] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Associative alignment for few-shot image classification. In *Proceedings of European Conference on Computer Vision*, pages 18–35, Glasgow, UK, November 2020. Springer.

[Afrasiyabi *et al.*, 2022] Arman Afrasiyabi, Hugo Larochelle, Jean-François Lalonde, and Christian Gagné. Matching feature sets for few-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9014–9024, New Orleans, USA, June 2022. IEEE.

[Agarwal *et al.*, 2021] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, , Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Proceedings of 34th Annual Conference on Neural Information Processing Systems*, pages 4078–4088, 4699–4711, December 2021. Neural Information Processing Systems Foundation.

[Baik *et al.*, 2021] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9465–9474, Nashville, USA, June 2021. IEEE.

[Bendou *et al.*, 2022] Yassir Bendou, Yuqing Hu, Raphael Lafargue, Giulia Lioi, Stéphane Pateux, and Vincent Gripon. Easy-ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components. *Journal of Imaging*, 8(7):179, 2022.

[Bertinetto *et al.*, 2019] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *Proceedings of 7th International Conference on Learning Representations*, New Orleans, USA, May 2019. International Conference on Learning Representations.

[Bouniot *et al.*, 2022] Quentin Bouniot, Ievgen Redko, Romaric Audigier, Angélique Loesch, and Amaury Habrard. Improving few-shot learning through multi-task representation learning theory. In *Proceedings of European Conference on Computer Vision*, pages 435–452, Tel Aviv, Israel, October 2022. Springer.

[Chen *et al.*, 2019] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *Proceedings of 7th International Conference on Learning Representations*, New Orleans, USA, May 2019. International Conference on Learning Representations.

[Cui *et al.*, 2017] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2921–2930, Venice, Italy, February 2017. IEEE.

[Dvornik *et al.*, 2019] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3723–3731, Seoul, Korea, February 2019. IEEE.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of International Conference on Machine Learning*, pages 1126–1135, Sydney,Australia, July 2017.

[Gidaris *et al.*, 2019] Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8059–8068, Seoul, Korea, February 2019. IEEE.

[Horváth *et al.*, 2021] Miklós Z Horváth, Mark Niklas Müller, Marc Fischer, and Martin Vechev. Boosting randomized smoothing with variance reduced classifiers. *arXiv preprint arXiv:2106.06946*, 2021.

[Ionescu *et al.*, 2015] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2965–2973, Santiago, Chile, February 2015. IEEE.

[Kang *et al.*, 2021] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8822–8833, Nashville, USA, June 2021. IEEE.

[Kifer *et al.*, 2004] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the 31st International Conference on Very Large Databases*, pages 180–191, Toronto, Canada;, September 2004. Morgan Kaufmann.

[Liu *et al.*, 2020] Bin Liu, Yue Cao, Yutong Lin, , Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Proceedings of European Conference on Computer Vision*, pages 438–455, Glasgow, UK, November 2020. Springer.

[Ma *et al.*, 2019] Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, and Wael Abd-Almageed. Partner-assisted learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10573–10582, Seoul, Korea, February 2019. IEEE.

[Ma *et al.*, 2022a] Rongkai Ma, Pengfei Fang, Gil Avraham, Yan Zuo, Tianyu Zhu, Tom Drummond, and Mehrtash Harandi. Learning instance and task-aware dynamic kernels for few-shot learning. In *Proceedings of European Conference on Computer Vision"*, pages 257–274. Springer, 2022.

[Ma *et al.*, 2022b] Rongkai Ma, Pengfei Fang, Tom Drummond, and Mehrtash Harandi. Adaptive poincaré point to set distance for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1926–1934, Austin, Texas, August 2022. AAAI.

[Mangla *et al.*, 2020] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, , and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2218–2227, Snowmass, USA, March 2020. IEEE.

[Rajasegaran *et al.*, 2020] Jathushan Rajasegaran, Salman Khan, , Fahad Shahbaz Khan, and Mubarak Shah. Self-supervised knowledge distillation for few-shot learning. *arXiv preprint arXiv:2006.09785*, 2020.

[Scott *et al.*, 2021] Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. von mises-fisher loss: An exploration of embedding geometries for supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10612–10622, Nashville, USA, June 2021. IEEE.

[Shiryaev, 2016] Albert N Shiryaev. *Probability-1*, volume 95. Springer, 2016.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of 31st Annual Conference on Neural Information Processing Systems*, pages 4078–4088, Long Beach, USA, December 2017. Neural Information Processing Systems Foundation.

[Sung *et al.*, 2021] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10836–10846, Nashville, USA, June 2021. IEEE.

[Tian *et al.*, 2020] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of European Conference on Computer Vision*, pages 266–282, Glasgow, UK, November 2020. Springer.

[Tseng *et al.*, 2020] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*, 2020.

[Wang *et al.*, 2020] Zeyuan Wang, Yifan Zhao, Jia Li, and Yonghong Tian. Cooperative bi-path metric for few-shot learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1524–1532, Seattle, USA, October 2020. ACM.

[Wertheimer *et al.*, 2021] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8012–8021, Nashville, USA, 2021.

[Wu *et al.*, 2021] Jiamin Wu, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Task-aware part mining network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8433–8442, Nashville, USA, June 2021. IEEE.

[Xie *et al.*, 2022] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, New Orleans, USA, June 2022. IEEE.

[Xu *et al.*, 2021a] Chengming Xu, Yanwei Fu, Chen Liu, Chengjie Wang, Jilin Li, Feiyue Huang, Li Zhang, and Xiangyang Xue. Learning dynamic alignment via meta-filter for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5182–5191, Nashville, USA, June 2021. IEEE.

[Xu *et al.*, 2021b] Jing Xu, Xinglin Pan, Xu Luo, Wenjie Pei, and Zenglin Xu. Exploring category-correlated feature for few-shot image classification. *arXiv preprint arXiv:2112.07224*, 2021.

[Yang *et al.*, 2013] Jing Yang, Xiaoqin Zeng, Shuiming Zhong, and Shengli Wu. Effective neural network ensemble approach for improving generalization performance. *IEEE transactions on neural networks and learning systems*, 24(6):878–887, 2013.

[Yang *et al.*, 2021] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: distribution calibration. In *Proceedings of 9th International Conference on Learning Representations*, New Orleans, USA, May 2021. International Conference on Learning Representations.

[Yu *et al.*, 2022] Tianyuan Yu, Sen He, Yi-Zhe Song, and Tao Xiang. Hybrid graph neural networks for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3179–3187, Austin, Texas, 2022.

[Zhang *et al.*, 2020a] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12203–12213, Seattle, USA, June 2020. IEEE.

[Zhang *et al.*, 2020b] Manli Zhang, Jianhong Zhang, , and Songfang Huang. Iept: Instance-level and episode-level pretext tasks for few-shot learning. In *Proceedings of 7th International Conference on Learning Representations*, Addis Ababa, Ethiopian Empire, May 2020. International Conference on Learning Representations.

[Zhou *et al.*, 2021] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8402–8411, Seoul, Korea, February 2021. IEEE.