

Article

Enhancing Few-Shot Learning in Lightweight Models via Dual-Faceted Knowledge Distillation

Bojun Zhou ¹, Tianyu Cheng ², Jiahao Zhao ¹, Chunkai Yan ¹, Ling Jiang ¹, Xinsong Zhang ² and Juping Gu ^{1,3,*}¹ School of Information Science and Technology, Nantong University, Nantong 226019, China; zhoubj@ntu.edu.cn (B.Z.)² School of Electrical Engineering, Nantong University, Nantong 226019, China³ School of Electronic & Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China

* Correspondence: gu.jp@ntu.edu.cn

Abstract: In recent computer vision research, the pursuit of improved classification performance often leads to the adoption of complex, large-scale models. However, the actual deployment of such extensive models poses significant challenges in environments constrained by limited computing power and storage capacity. Consequently, this study is dedicated to addressing these challenges by focusing on innovative methods that enhance the classification performance of lightweight models. We propose a novel method to compress the knowledge learned by a large model into a lightweight one so that the latter can also achieve good performance in few-shot classification tasks. Specifically, we propose a dual-faceted knowledge distillation strategy that combines output-based and intermediate feature-based methods. The output-based method concentrates on distilling knowledge related to base class labels, while the intermediate feature-based approach, augmented by feature error distribution calibration, tackles the potential non-Gaussian nature of feature deviations, thereby boosting the effectiveness of knowledge transfer. Experiments conducted on MiniImageNet, CIFAR-FS, and CUB datasets demonstrate the superior performance of our method over state-of-the-art lightweight models, particularly in five-way one-shot and five-way five-shot tasks.



Citation: Zhou, B.; Cheng, T.; Zhao, J.; Yan, C.; Jiang, L.; Zhang, X.; Gu, J. Enhancing Few-Shot Learning in Lightweight Models via Dual-Faceted Knowledge Distillation. *Sensors* **2024**, *24*, 1815. <https://doi.org/10.3390/s24061815>

Academic Editor: Loris Nanni

Received: 31 January 2024

Revised: 8 March 2024

Accepted: 10 March 2024

Published: 12 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: few-shot classification; knowledge distillation; model compression; distribution calibration

1. Introduction

In today's digital era, computer vision, as a key branch of AI, is developing rapidly and playing an increasingly important role in many application fields. Especially in the task of image classification, high-performance vision processing systems are crucial to many industries, such as autonomous driving, medical diagnosis, and power inspection [1–3]. Recently, researchers have generally adopted complex and large-scale deep learning models in order to improve the classification performance of these systems [4–6]. Although these models are designed to have excellent performance, they face a series of challenges when actually deployed in computing resource-constrained environments, such as high computing costs and limited storage space [7].

In view of these limitations, how to improve the performance of lightweight models for classification tasks has become a key research topic in the field of machine learning. Few-shot learning (FSL) is designed to enable models to learn from limited data samples and make accurate predictions [8]. When FSL is applied in some specific scenarios, such as electric mobile inspection or small UAVs, lightweight models are more suitable. However, training lightweight models like large models often encounters the problem of overfitting, causing the actual performance of the model to be far less than expected. To this end, the main goal of this study is to explore a novel method to improve the performance of lightweight models in few-shot classification (FSC) tasks. With this approach, we hope to narrow the performance gap between lightweight and large models on FSC tasks.

In the field of FSC, Conv4 and ResNet12 represent prevalent neural network architectures [9,10], each offering distinct levels of complexity. Compared to Conv4, ResNet12 is a more complex or larger model. The choice of which model to use may depend on the needs of the particular task as well as the available computational resources. After an in-depth study of the performance differences between a lightweight model (Conv4) and a complex model (ResNet12) in the FSC task, our analysis reveals significant differences in two main areas. First, as shown in Figure 1a, we compare the accuracies of Conv4 and Resnet12 in classifying the base class samples in the pre-training phase. This comparison clearly shows that there is a considerable difference between the two in terms of classification accuracy. The lightweight model does not seem to be able to extract enough information from the input data of the base class to achieve a similar level of performance as the complex model. Second, by visualizing the output features of Conv4 and Resnet12 on the test images, as illustrated in Figure 1b, we further observe that the two models also exhibit significant differences in feature representation. This difference points to a fundamental difference in feature extraction and representation capabilities, providing a key clue to understanding why the lightweight model lags behind in performance.

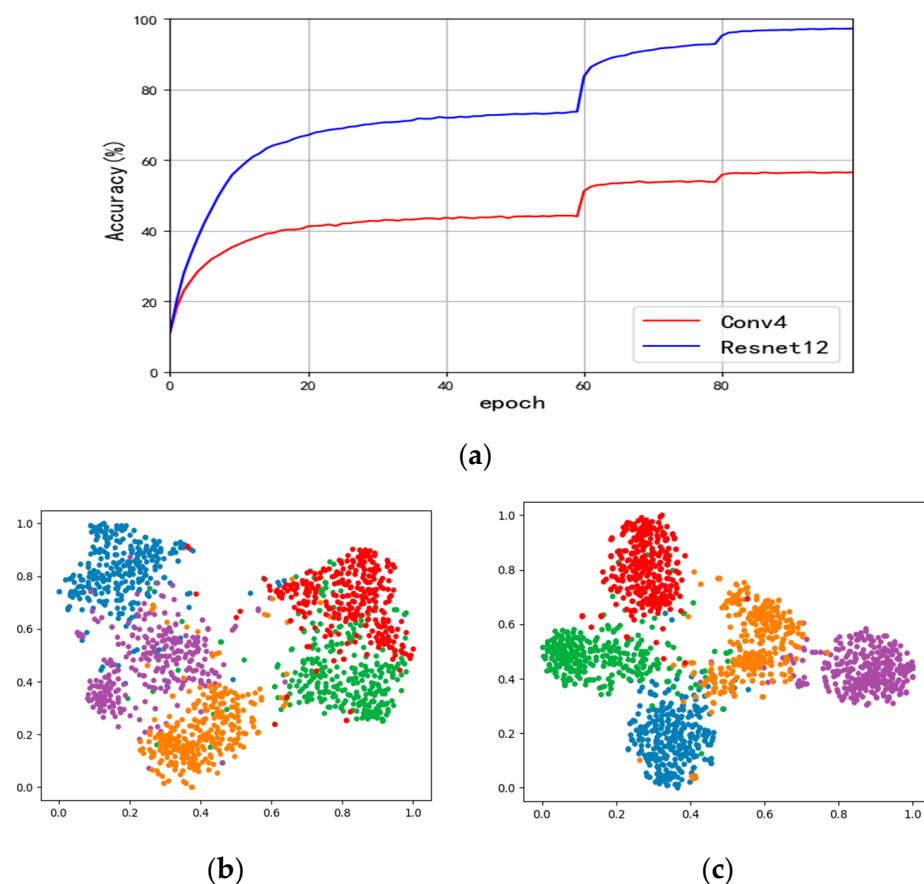


Figure 1. Performance difference between lightweight model (Conv4) and complex model (ResNet12) in few-shot classification tasks. (a) The accuracy of classifying base class samples; (b) the feature distributions of 5 randomly selected classes post Conv4 processing; (c) the feature distributions of 5 randomly selected classes post ResNet12 processing, where each color represents a distinct category of features.

Based on the above analysis, this study considers lightweight and complex models as student and teacher models, respectively, and conducts a model compression task using a knowledge distillation-based approach, aiming to enhance the performance of lightweight models in FSC tasks. Specifically, this study first applies an output-based knowledge distillation algorithm to enhance the lightweight model's ability to classify the base class,

with the expectation that the student model can mimic and learn the output knowledge of the teacher model. In addition, to further enhance the feature representation capability of the lightweight model, we adapt a distillation method based on feature error distribution calibration. This method works by adjusting the intermediate features of the student model to make it closer to the feature output of the teacher model.

As shown in Figure 2a, previous intermediate feature-based knowledge distillation methods typically take the feature output of the teacher model as the desired output and approximate the teacher model by minimizing the mean square error function of the features. However, the effectiveness of this approach depends on the assumption that the data follow a Gaussian distribution, as the mean square error loss function is a natural choice for maximum likelihood estimation under a Gaussian distribution (proof is provided in Section 3.3.1). If the feature errors do not follow a Gaussian distribution, relying only on the mean square error may result in the training of the student model deviating from the optimum. With this in mind, this study proposes a novel knowledge distillation strategy. As shown in Figure 2b, this strategy first applies Gaussian calibration to the distribution of feature errors between the teacher and student models, and then constructs the mean square error loss function. This approach helps the student model mimic the intermediate features of the teacher model more accurately, thus improving its classification performance.

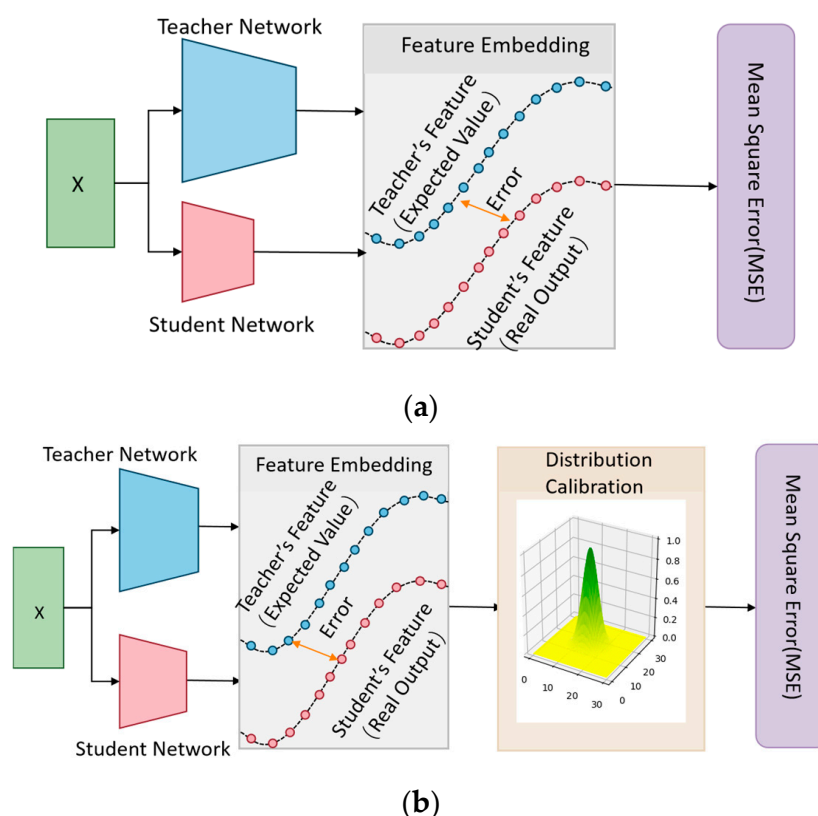


Figure 2. A novel intermediate feature-based knowledge distillation method. (a) Previous intermediate feature-based knowledge distillation method; (b) a knowledge distillation method based on feature error distribution calibration proposed in this study.

In summary, the contributions of this paper are as follows:

- (1) We investigate the application of the dual-faceted knowledge distillation method to the task of few-shot model compression to enhance the performance of lightweight models.
- (2) We develop a novel method for calibrating feature error distribution that significantly enhances the performance of feature-based knowledge distillation. Furthermore, we

provide theoretical proof to substantiate our proposed method, offering valuable insights and establishing a robust framework for future research in this field.

- (3) We demonstrate the effectiveness of our proposed method by validating it on three benchmark datasets. Our proposed method outperforms all other methods and achieves the best performance.

2. Related Work

2.1. Few-Shot Classification

Depending on the learning paradigm used to acquire prior knowledge, most of the current few-shot classification methods can be categorized into meta-learning-based methods and transfer learning-based methods [11]. Meta-learning, also known as learn to learn [12], is the dominant learning paradigm for solving few-shot image classification tasks. This learning strategy involves randomly selecting a number of few-shot classification tasks, using epoch training to acquire the meta-knowledge implicit in these tasks, and quickly processing unseen tasks based on the meta-knowledge [13]. Depending on the type of meta-knowledge learned, these methods are further classified into optimization-based methods and metric-based methods. Compared to the former, metric-based methods generally have better classification performance. Learning to obtain good features for distinguishing different classes of samples is the key to achieving good performance for these methods. For example, PN+rot [14] improved the generalization performance of metric models by adding the rotation prediction task from self-supervised learning to the meta-learning process. Recently, refs. [15,16] have made improvements in improving the generalization performance of prototype networks, with the former focusing on normalizing the prototypes computed based on the support samples in each sub-task, while the latter adopted the approach of estimating in regenerated Hilbert space to obtain the relative prototypes of each class of support samples and introducing the Tikhonov regularization technique in the training process.

Recent research has shown that in transfer learning, using a whole base class dataset to pre-train the model and fine-tune the classifier using a small number of samples can achieve comparable performance to meta-learning. In this context, early work [17] focused on pre-training convolutional neural networks using the cross-entropy loss function, providing a powerful baseline for the task of FSC. In this classification framework, how to obtain an effective feature representation in the pre-training phase is crucial to improving the performance of FSC. In order to enhance the generalization ability of the traditional cross-entropy loss function, Neg-Cosine [18] and S2M2 [19] introduced non-negative-spaced cosine loss function and streaming mixing techniques, respectively. These methods are proposed to enhance the classification accuracy by improving the loss function and enhancing the richness of the feature representation. In addition, self-supervised tasks have been shown to enhance the generalization performance of FSC models in a transfer learning framework. For instance, IE [20] augmented the model by integrating self-supervised tasks such as rotation prediction and instance discrimination, thereby enhancing the model's invariance and equivariance to variations in input. Meanwhile, CCF [21] introduced the task of generating features during the pre-training process to further enhance the model's representational capabilities.

Considering recent advances in the task of FSC, a common trend can be observed: whether based on meta-learning or transfer-learning approaches, scholars generally tend to adopt complex models to enhance the capability of feature representation. However, these complexity models tend to consume a large number of computational resources in practical applications and also lead to higher time delays. These shortcomings limit the application of these methods in resource-constrained real-world scenarios requiring fast responses. In view of this, this paper proposes a few-shot model compression method based on a dual-knowledge distillation strategy that has important practical significance and research value.

2.2. Knowledge Distillation

Knowledge distillation is a very effective method of model compression that employs a high-capacity teacher model to instruct a compact student model so that the student has comparable performance to the former. The inter-model knowledge distillation method was first proposed by Hinton [22]. In order to better utilize the knowledge information contained in the teacher's model, recent work has focused more on how to better extract the feature knowledge implicit in the middle layer of the network. For example, FitNet [23] directly minimizes the L2 paradigm between the feature outputs of the corresponding layers between the student network and the teacher network. AT [24] proposes to use the spatial attention of the features in the implicit layer of the teacher network as the knowledge and instructs students to mimic the teacher's attentional feature maps. SP [25] and RKD [26] compute the relationship matrix between image features in the student model and the teacher model, enabling the student to better learn the teacher's relational knowledge. PKT [27] and NST [28] align the overall distribution of the student and teacher output features using the mutual information scatter measure and maximum mean difference, respectively.

Several recent research works have also introduced knowledge distillation techniques to FSC tasks. For example, RFS [29] and SKD [30] utilize a network regeneration strategy to obtain a more robust feature extraction model through continuous evolution between multiple generations. PAL [31] expects the student model to align to the teacher model in terms of output logical values and implicit features, respectively. BML [32] utilizes the complementary nature of meta-learning and transfer learning to train the student model separately while using deep mutual learning techniques to facilitate collaborative learning among student models during the training process. In the above work, both the student and the teacher model are self-distilled using complex models with the same structure. Contrary to the aforementioned approaches, this study employs knowledge distillation as a strategy for transferring knowledge from a complex model to a lightweight counterpart, with the primary goal of enhancing the performance of the lightweight model.

3. Main Approach

3.1. Overview

According to the process setup of the few-shot image classification task [33], the given image dataset is randomly partitioned into three subsets, i.e., the base class dataset D_b , the validation dataset D_{val} , and the novel class dataset D_n . These three subsets, which have different class labels, are used for the training of the model, the validation of the model, and the few-shot testing, respectively. In the model validation and few-shot testing phases, several N-way K-shot classification tasks need to be constructed, each of which contains a collection of support samples and a collection of query samples. K samples from N categories are randomly selected from D_{val} or D_n to construct the support set $D_s = \{(x_k, y_k)\}_{k=1}^{NK}$, where x_k denotes the k -th image and y_k is its corresponding category label. The remaining samples from N categories form the query set, denoted as $D_q = \{(x_q, y_q)\}_{q=1}^Q$, where x_q denotes the q -th query sample.

As shown in Figure 3, the methodological framework proposed in this paper is divided into three main phases. In the first phase, we train a well-performing teacher network (Resnet12) on D_b using a weighted sum of cross-entropy loss and self-supervised loss. This step aims to obtain robust feature representations by combining standard supervised learning and self-supervised learning. The second stage involves the use of an innovative dual-knowledge distillation strategy that aims to efficiently migrate the output knowledge and intermediate layer feature knowledge from the teacher model to the student model (Conv4).

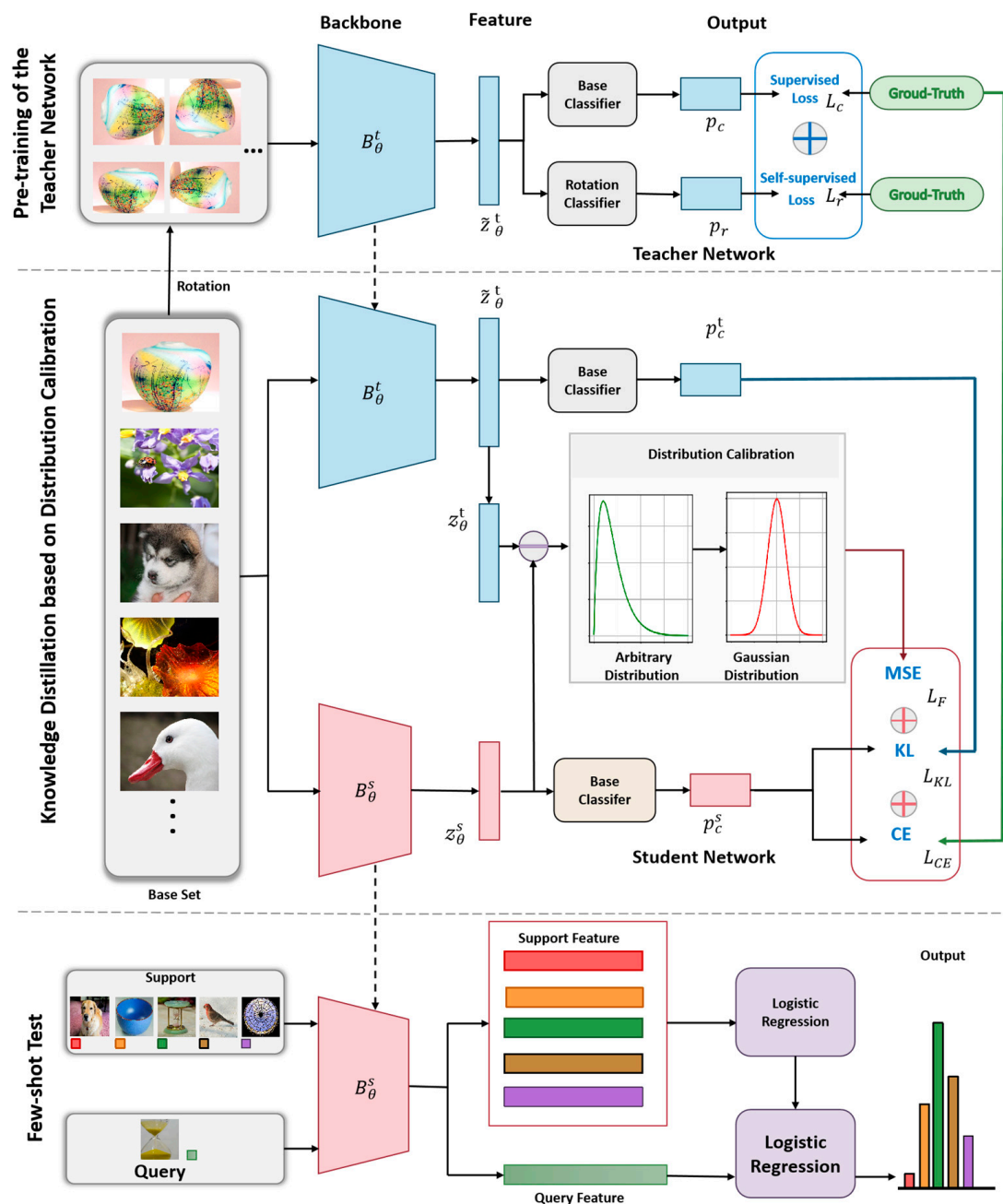


Figure 3. The framework diagram of our method. It presents a framework for model compression in few-shot image classification using dual-faceted knowledge distillation. Initially, a teacher network is trained with combined supervised and self-supervised losses. Knowledge is then distilled to a lightweight student model through two strategies focusing on output and feature knowledge. Finally, the student model is evaluated on FSC tasks, demonstrating its ability to generalize effectively with limited data.

The backbone of our proposed model consists of a four-layer convolutional network, Conv4. Each layer comprises a sequence of a convolutional layer, batch normalization, and a ReLU activation function. The first three layers are followed by a max-pooling layer to reduce the dimensionality of the features. Contrary to the preceding layers, the final convolutional layer is directly followed by an adaptive average pooling layer, which reduces the spatial dimensions of the feature maps to 1×1 , facilitating the extraction of the feature vectors.

In this way, we accomplish the task of compressing the few-shot image classification model, aiming to reduce the complexity of the model while minimizing the performance gap between the two. In the third phase, we deploy the trained lightweight student network on D_n for a few-shot image classification test. This phase mainly verifies the effectiveness of the lightweight model after model compression on the new task. The key steps and detailed flow of this method will be described and discussed in detail in the remainder of this section.

3.2. Teacher Model Construction

Recent studies have shown that self-supervised learning effectively improves the generalization performance of transfer-learning models in FSC tasks [34]. Based on this, in this paper, we use a linear combination of cross-entropy and self-supervised loss function to pre-train the teacher network model and select rotation prediction as the self-supervised learning task. The structure of the teacher model mainly consists of a backbone network B_{θ_t} , two linear classifiers, and their corresponding Softmax classification layers, where θ_t represents the parameters of the backbone network. A linear classifier $C_b(\cdot)$ is used for the prediction of the base classes with a parameter matrix of $W_b = \{w_c\}_{c=1}^C$. Another linear classifier $C_r(\cdot)$ is used for the prediction of the rotated classes, and its parameter matrix is $W_r = \{w_r\}_{r=1}^R$. $w_c, w_r \in R^d$ are d -dimensional weight vectors, and C and R are denoted as the number of base categories and rotated categories, respectively.

In the base class dataset, a number of batches of L images are randomly selected, and each image is rotated by 0° , 90° , 180° , and 270° and labeled with its rotation class $o_i = [0, 1, 2, 3]$. Each image is inputted into the teacher network, and the backbone of the network maps it into the d -dimensional feature space. The feature can be represented as:

$$z_i^t = B_{\theta_t}(x_i) \in \mathcal{R}^d \quad (1)$$

Subsequently, the probabilistic predicted values p_{ic} and p_{ir} are obtained for the base class and the rotated class through the two linear classifiers and the corresponding Softmax layers, respectively, which are computed as follows:

$$\begin{aligned} p_{ic} &= \frac{e^{w_c z_i^t}}{\sum_{c=1}^C e^{w_c z_i^t}} \\ p_{ir} &= \frac{e^{w_r z_i^t}}{\sum_{r=1}^R e^{w_r z_i^t}} \end{aligned} \quad (2)$$

Then, the cross-entropy loss function between the base class predicted values and the rotated predicted values and their corresponding ground truth values are calculated as:

$$\begin{aligned} L_c(\theta_t, W_b) &= \sum_{i=1}^{RL} \sum_{c=1}^C y_{ic} \log p_{ij} \\ L_r(\theta_t, W_r) &= \sum_{i=1}^{RL} \sum_{r=1}^R o_{ir} \log p_{ir} \end{aligned} \quad (3)$$

The total loss function for pre-training the teacher network is denoted as:

$$L_t(\theta_t, W_b, W_r) = L_c(\theta_t, W_b) + \lambda L_r(\theta_t, W_r), \quad (4)$$

where λ is the adjustable weight parameter. Utilizing Equation (4), the pre-training of the teacher network is accomplished through the application of the gradient descent algorithm.

3.3. Dual-Faceted Knowledge Distillation Strategy

3.3.1. Theoretical Foundation

Since the student model has a very small capacity to handle complex self-supervised tasks, the student model mainly learns the base class classification task from the teacher

model. The network structure of the student model mainly consists of a backbone B_{θ_s} , a linear classifier, and a corresponding Softmax classification layer, where θ_s represents the parameters of the backbone of the model. The linear classifier $C_s(\cdot)$ is used for the prediction of the base classes, and its parameter matrix is $W_s = \{w_c^s\}_{c=1}^C$, where $w_c^s \in R^m$ is an m -dimensional classification weight vector.

A number L of image batches are randomly selected from the base class data set, where the corresponding base class label of the i -th image x_i is y_i . Each image is input to both the teacher network and the student network. The teacher's and student's backbone networks are mapped to d - and m -dimensional feature spaces, respectively, to obtain features represented as follows:

$$\begin{aligned} \hat{z}_i^t &= B_{\theta_t}(x_i) \in \mathcal{R}^d \\ z_i^s &= B_{\theta_s}(x_i) \in \mathcal{R}^m \end{aligned} \quad (5)$$

Since the dimensions between the output features of the student and the teacher network are not the same, direct matching cannot be performed. For this reason, a linear matching layer $P(\cdot)$ is introduced, which maps the feature output of the teacher network to the m -dimensional feature space. At this time, the features of the teacher network are represented as follows:

$$z_i^t = P(\hat{z}_i^t) \in \mathcal{R}^m. \quad (6)$$

At this point, the error between the transformed teacher network features and the student network features is computed to obtain the m -dimensional feature error variable e , whose i -th observation is denoted as:

$$e_i = z_i^t - z_i^s. \quad (7)$$

Proposition 1. *For a given image x , suppose that the intermediate feature output variable of the student network is z^s with L observations $\{z_i^s\}_{i=1}^L$ and the intermediate feature output variable of the teacher network is z^t with L observations $\{z_i^t\}_{i=1}^L$. The intermediate feature-based knowledge distillation process takes the feature output value z^t of the teacher network as the target variable, and the student network makes the feature output value z^s approximate to the target variable by optimizing the model parameters. When the error variable e between the output features of the teacher network and the student network obeys a Gaussian distribution, the mean square error function is the optimal loss function.*

Proof of Proposition 1. The knowledge distillation process for intermediate features can be described using the joint probability density function $p(x, e)$. By applying conditional probability distribution, this function can be further expressed as:

$$p(x, e) = p(e|x)p(x). \quad (8)$$

For a set of L independently and identically distributed training samples, the likelihood function is:

$$L(\theta_s) = \prod_{i=1}^L p(e_i, x_i) = \prod_{i=1}^L p(e_i|x_i)p(x_i). \quad (9)$$

Taking the logarithm of both sides of Equation (9) yields:

$$\ln L(\theta_s) = \sum_{i=1}^L \ln p(e_i|x_i) + \sum_{i=1}^L \ln p(x_i). \quad (10)$$

To maximize the likelihood function, the objective function $L_F(\theta_s)$ is defined as:

$$L_F(\theta_s) = -\ln L(\theta_s) = -\sum_{i=1}^L \ln p(e_i|x_i) - \sum_{i=1}^L \ln p(x_i). \quad (11)$$

In Equation (11), the second term is independent of the student network and the objective function is simplified to:

$$L_F(\theta_s) = -\sum_{i=1}^L \ln p(e_i|x_i). \quad (12)$$

Assuming that the errors $e \in R^m$ follow a Gaussian distribution with mean 0 and variance σ^2 , $p(e|x)$ is determined as:

$$p(e|x) = \frac{1}{(2\pi)^{m/2}\sigma^m} \exp\left\{-\frac{e^T e}{2\sigma^2}\right\}. \quad (13)$$

Substituting Equation (13) into the objective Function (12), we get:

$$\begin{aligned} L_F(\theta_s) &= -\sum_{i=1}^L \ln \frac{1}{(2\pi)^{m/2}\sigma^m} \exp\left\{-\frac{e_i^T e_i}{2\sigma^2}\right\} \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^L e_i^T e_i + \frac{Lm}{2} \ln 2\pi + Lm \ln \sigma \end{aligned} \quad (14)$$

In Equation (14), the second and third terms are independent of the student network, and since the covariance σ^2 is a constant, the objective function can be further simplified to:

$$L_F(\theta_s) = \frac{1}{2} \sum_{i=1}^L e_i^T e_i. \quad (15)$$

This demonstrates that under the assumption of Gaussian-distributed errors, the optimal objective function for knowledge distillation of intermediate features is the mean squared error of those features. $\square$

3.3.2. Intermediate Feature-Based Distillation

According to Proposition 1, it can be seen that in intermediate feature-based knowledge distillation, the model can achieve the optimum by using the mean square error function as the objective function only when the error variable between the output features of the teacher model and the student model obeys a Gaussian distribution. However, in practice, the probability density distribution of the feature errors between the teacher model and the student model does not necessarily satisfy the Gaussian distribution, and when the mean square error function is used to establish the objective function, the result will deviate from the true optimum.

In order to solve the problem, this paper proposes a novel knowledge distillation algorithm based on feature error distribution calibration. Firstly, the probability density distribution of the feature error variable between the teacher model features and the student model features is calibrated to a Gaussian distribution. In this paper, Tukey Ladder of Powers transformation [35] is used to correct the feature error variables in equalization (16) and scale to unit length, which is calculated as follows:

$$\tilde{e}_i = \begin{cases} \frac{(e_i + \epsilon)^\beta}{\|(e_i + \epsilon)^\beta\|_2}, & \beta \neq 0 \\ \frac{\log(e_i + \epsilon)}{\|(e_i + \epsilon)^\beta\|_2}, & \beta = 0 \end{cases} \quad (16)$$

where ε is a very small positive real number and β is an adjustable parameter, the skewness of the distribution can be controlled by adjusting β during the experimental process, and the distribution of the feature error variables can be made close to the Gaussian distribution.

After distribution calibration, the mean square error loss function based on the intermediate features is calculated as:

$$L_F(\theta_s) = \frac{1}{2} \sum_{i=1}^L \tilde{e}_i^T \tilde{e}_i. \quad (17)$$

3.3.3. Output-Based Distillation

For the features z^t and z^s obtained from the teacher and student models, respectively, they are each passed through their respective base class predictive linear classifiers followed by the corresponding Softmax layers. This process yields the probability outputs of the base classes for the student network, calculated as follows:

$$p_{ic}^s = \frac{e^{w_c^s z_i^s}}{\sum_{c=1}^C e^{w_c^s z_i^s}}. \quad (18)$$

At the same time, the soft target for the probability output value of the base classes in the teacher model is obtained as:

$$p_{ic}^t = \frac{e^{\tau w_c^t z_i^t}}{\sum_{c=1}^C e^{\tau w_c^t z_i^t}}, \quad (19)$$

where τ represents the temperature parameter. Consequently, we employ the classical knowledge distillation algorithm, enabling the student model to learn the soft target output by the teacher network. The calculation formula for the Kullback–Leibler (KL) divergence between the soft target output by the teacher model and the output information of the student model is as follows:

$$L_{KL}(\theta_s, W_s) = - \sum_{i=1}^L \sum_{c=1}^C p_{ic}^s \log \frac{p_{ic}^s}{p_{ic}^t}. \quad (20)$$

3.4. Model Evaluation and Testing

To enhance the base class prediction capabilities of the student model, this study also introduces the cross-entropy loss function to reduce the deviation between the predicted values and ground truth. The calculation formula for the cross-entropy loss function is:

$$L_{CE}(\theta_s, W_s) = - \sum_{i=1}^L \sum_{c=1}^C y_{ic} p_{ic}^s. \quad (21)$$

As a result, the overall loss function for knowledge distillation in the student model is calculated as:

$$L_s(\theta_s, W_s) = L_{CE}(\theta_s, W_s) + \alpha_1 L_{KL}(\theta_s, W_s) + \alpha_2 L_F(\theta_s), \quad (22)$$

where α_1 and α_2 are balancing coefficients, controlling the importance of each loss term in the loss function. Based on Equation (22), the optimization of parameters in the student model is completed using the gradient descent algorithm.

After the model training is completed, the base class classifier in the student network is removed, and the parameters θ_s in the backbone network B_{θ_s} are fixed. The features of images from the support set and the query set are extracted by feeding them into the student network. The feature representations of the k -th support sample and the q -th query sample are as follows:

$$\begin{aligned} z_k &= B_{\theta_s}(x_k) \in \mathcal{R}^t \\ z_q &= B_{\theta_s}(x_q) \in \mathcal{R}^t \end{aligned} \quad (23)$$

Using z_k and the corresponding label y_k , a logistic regression classifier $g_{\varnothing}(\cdot)$ with parameter $\varnothing$ is trained. The classification prediction for the q -th query sample is represented as:

$$\hat{y}_q = g_{\varnothing}(z_q). \quad (24)$$

As detailed in Section 3, the implementation process of this method is displayed in Algorithm 1.

Algorithm 1 Implementation: Few-Shot Model Compression Algorithm.

Input: Base class dataset D_b , validation dataset D_{val} , and the novel class dataset D_n , the teacher network $\{B_{\theta}^t(\cdot), C_b(\cdot), C_r(\cdot)\}$, the student network $\{B_{\theta}^s(\cdot), C_s(\cdot)\}$, temperature parameter τ , hyperparameter $\alpha_1, \alpha_2, \beta$.

Output: The predicted value $\hat{y}_q$ of query samples in D_n

Stage 1: Teacher network pre-training

While epoch $\leq$ maximum number of the iteration

A batch of images is randomly selected from D_b .

Images are fed into the backbone of the teacher network to extract the feature.

Obtain the base class and rotation class probability values.

Pre-train the teacher network according to Equation (4).

Stage 2: Few-shot model compression

While epoch $\leq$ maximum number of the iteration

A batch of images is randomly selected from D_b .

The image is separately fed into the backbone of the teacher and the student networks to extract features.

Obtain the base class probability values from the teacher network and the student network, respectively.

Calibrate the feature error distribution between the student network and the teacher network according to Equation (16).

Calculate the knowledge distillation loss function for intermediate features according to Equation (17).

Calculate the KL divergence-based loss function between the predicted output values of the student network and the teacher network according to Equation (20).

Calculate the cross-entropy loss function of the student network according to Equation (21).

Train the student network according to Equation (22).

Stage 3: Few-shot model testing

While epoch $\leq$ maximum number of the iteration

Images from D_n are processed through the feature extractor to obtain the feature.

Train classifier $g_{\varnothing}(\cdot)$ for the novel classes.

Test on the query set from D_n .

4. Experiments

4.1. Dataset

In this study, we conduct comprehensive experiments to evaluate the performance of our proposed method on three benchmark databases for few-shot image classification. MiniImageNet [36] is a subset derived from the larger ImageNet dataset. CIFAR-FS [37] is constructed based on the CIFAR100 dataset. Caltech-UCSD Bird-2002011 (CUB) [38] is the most widely used benchmark in the domain of fine-grained image classification.

4.2. Experimental Setup

In this study, all experiments were implemented on a Dell workstation equipped with an NVIDIA 3090Ti GPU at Nantong University, Nantong, China. The software used for these experiments was Pytorch version 1.7.0. For the few-shot image classification tasks, the most commonly used complex model is ResNet12, which we adopted as the teacher model in our method. The most frequently utilized lightweight model in this domain is the four-layer convolutional neural network (Conv4), selected as the student model in our approach. Each convolutional layer of this model contains 64 filters. During the training phase, the stochastic gradient descent (SGD) optimizer was employed, with the momentum

set to 0.9 and weight decay set to 5×10^{-4} . The models were trained for 160 epochs, with an initial learning rate of 0.025. The learning rate was reduced to one-tenth of its value every 30 epochs after the first 70 epochs. For the calculation of the soft targets in the teacher network's output, the temperature coefficient was set to 0.25. During the testing phase, experiments were conducted in both five-way one-shot and five-way five-shot tasks. For this purpose, 2000 random classification subtasks were generated in the novel class dataset. In each subtask, 15 images per class were randomly selected as query images for testing. The performance was evaluated based on the average accuracy across all subtasks, with the standard deviation of the accuracy given under a 95% confidence interval.

4.3. Methodological Validation

4.3.1. Model Compression Efficacy

In pursuit of enhancing the few-shot image classification capabilities of lightweight models, this study introduces a novel model compression method employing dual-faceted knowledge distillation. To validate the effectiveness of this innovative approach, we conducted a series of evaluations comparing the performance of standard lightweight models (referred to as “baseline”), trained solely using cross-entropy loss, with models augmented by our novel compression technique (referred to as “model compression”, or MC). Table 1 presents the comparative classification accuracy of both the baseline and the MC strategy in five-way one-shot and five-way five-shot classification tasks on three datasets: MiniImageNet, CIFAR-FS, and CUB.

Table 1. Comparative classification accuracy (%) of baseline and our proposed method.

Method	Backbone	MiniImageNet		CIFAR-FS		CUB	
		1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
Baseline	Conv4-64	53.36	70.51	63.83	78.14	64.77	79.17
MC	Conv4-64	57.99	73.30	67.02	80.98	69.92	85.27

The results presented in Table 1 indicate that our model compression method significantly outperforms the baseline in classification accuracy for both five-way one-shot and five-way five-shot tasks across all benchmark datasets. For MiniImageNet, the MC method improves upon the baseline by 4.63 percentage points in the one-shot task and by 1.04 percentage points in the five-shot task. In the CIFAR-FS dataset, the performance gains are 3.19 percentage points for the one-shot and 2.84 percentage points for the five-shot tasks. Likewise, on CUB, MC records increases of 5.15 percentage points for the one-shot and 6.1 percentage points for the five-shot tasks. These results, further detailed in Section 3, underscore the effectiveness of our proposed compression method in enhancing the few-shot classification capabilities of lightweight models.

4.3.2. Feature Error Distribution Calibration

In this research, we introduce a novel knowledge distillation algorithm enhanced by feature error distribution calibration. This section examines the impact of the calibration phase on the algorithm's performance. We refer to the algorithm without calibration as KD-Plain and the one with calibration as KD-Improved. Table 2 illustrates the accuracy of both KD-Plain and KD-Improved on the MiniImageNet, CIFAR-FS, and CUB datasets for five-way one-shot and five-way five-shot tasks. The results, as shown in Table 2, consistently indicate that KD-Improved outperforms KD-Plain across all datasets and tasks.

Table 2. Comparative classification accuracy (%) of KD-Plain and KD-Improved methods.

Dataset	Backbone	Setting	KD-Plain	KD-Improved
MiniImageNet	Conv4-64	1-shot	57.01	57.99 (0.98 ↑)
	Conv4-64	5-shot	72.69	73.30 (0.61 ↑)
CIFAR-FS	Conv4-64	1-shot	65.98	67.02 (1.04 ↑)
	Conv4-64	5-shot	80.02	80.98 (0.96 ↑)
CUB	Conv4-64	1-shot	68.87	69.92 (1.05 ↑)
	Conv4-64	5-shot	84.24	85.27 (1.03 ↑)

The efficacy of our knowledge distillation algorithm, which includes the calibration step, is thus effectively demonstrated. This step addresses the common issue in distillation algorithms that rely on intermediate features, which is the assumption that feature deviations follow a Gaussian distribution. By calibrating these deviations towards a Gaussian model, our approach allows the mean squared error loss to work more effectively. Further details on the calibration process and its optimization criteria have been discussed in Section 3.3.

4.4. Comparative Studies

4.4.1. Comparison with Classical Knowledge Distillation Approaches

In this research, we benchmark our novel knowledge distillation algorithm, which employs feature error distribution calibration, against classical methods in the field within the context of model compression for few-shot image classification tasks. We categorized these methods into two groups: those utilizing output knowledge, such as Hinton’s knowledge distillation (HKD) [22], and those based on intermediate feature knowledge, such as FitNet [23], AT [24], SP [25], RKD [26], PKT [27], and NST [28]. The classification results of these methods are detailed in Table 3.

Table 3. Comparison of the proposed knowledge distillation method with classical approaches.

Method	Backbone	MiniImageNet		CIFAR-FS		CUB	
		1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
HKD [22]	Conv4-64F	57.33 ± 0.41	72.42 ± 0.34	66.27 ± 0.93	80.35 ± 0.69	68.28 ± 0.50	84.07 ± 0.28
FitNet [23]	Conv4-64F	57.43 ± 0.45	72.26 ± 0.34	66.40 ± 0.91	80.44 ± 0.67	68.45 ± 0.48	83.63 ± 0.29
AT [24]	Conv4-64F	57.86 ± 0.48	72.79 ± 0.35	66.75 ± 0.92	80.72 ± 0.68	67.90 ± 0.48	83.53 ± 0.30
SP [25]	Conv4-64F	57.26 ± 0.45	72.61 ± 0.35	66.72 ± 0.92	80.68 ± 0.69	68.37 ± 0.47	84.05 ± 0.29
RKD [26]	Conv4-64F	57.14 ± 0.45	72.89 ± 0.35	66.73 ± 0.91	80.70 ± 0.68	68.68 ± 0.47	84.03 ± 0.29
PKT [27]	Conv4-64F	57.61 ± 0.48	72.91 ± 0.31	66.84 ± 0.92	80.70 ± 0.67	68.13 ± 0.47	84.02 ± 0.29
NST [28]	Conv4-64F	57.82 ± 0.48	72.66 ± 0.35	66.83 ± 0.93	80.72 ± 0.68	68.07 ± 0.48	84.02 ± 0.29
Ours	Conv4-64F	57.99 ± 0.44	73.30 ± 0.36	67.02 ± 0.92	80.98 ± 0.69	69.92 ± 0.46	85.27 ± 0.28

Table 3 demonstrates that our method, which leverages both class intermediate feature knowledge and class output, consistently outperforms the traditional approaches on the MiniImageNet, CIFAR-FS, and CUB datasets. Notably, on CUB, our method shows a marked improvement over the best-performing traditional knowledge distillation technique, with an increase of 1.64% in one-shot and 1.2% in five-shot tasks. Our approach’s enhancements are also evident when directly compared to FitNet and AT, where it exhibits superior accuracy, surpassing AT by 2.02% in one-shot and by 1.74% in five-shot tasks on CUB. Furthermore, our algorithm outperforms SP and RKD, which do not account for the error distribution in relational features, as well as PKT and NST, which may neglect rich sample information with their simpler global relationship models. Our method achieves the largest performance gains over NST on CUB, with improvements of 1.85% in one-shot and of 1.25% in five-shot tasks. These results confirm the efficacy of our algorithm, particularly its feature error distribution correction component, in improving the few-shot image classification performance of lightweight models in compression scenarios.

4.4.2. Comparison with Other Methods

Our proposed method is compared with recent state-of-the-art (SOTA) lightweight model-based methods on MiniImageNet, CIFAR-FS, FC100, and CUB datasets. Consistent with the classification standards in related works, these methods are categorized into meta-learning and transfer learning paradigms. The comparative results of our method against these SOTA methods are detailed in Table 4.

Table 4. Comparison of classification results (%) on MiniImageNet, CIFAR-FS, and CUB.

Method	Backbone	MiniImageNet		CIFAR-FS		CUB	
		1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
Meta-learning paradigms							
MAML [12]	Conv4-64F	48.70 $\pm$ 1.75	63.11 $\pm$ 0.92	58.90 $\pm$ 1.90	71.50 $\pm$ 1.00	55.92 $\pm$ 0.95	72.09 $\pm$ 0.76
Prototypical [39]	Conv4-64F	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66	51.31 $\pm$ 0.91	70.77 $\pm$ 0.69	-	-
Relational [40]	Conv4-64F	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66	51.31 $\pm$ 0.91	70.77 $\pm$ 0.69
MetaOpt SVM [41]	Conv4-64F	52.87 $\pm$ 0.57	68.76 $\pm$ 0.48	-	-	62.45 $\pm$ 0.98	76.11 $\pm$ 0.69
PN+rot [14]	Conv4-64F	53.63 $\pm$ 0.43	71.70 $\pm$ 0.36	-	-	-	-
CovaMNet [42]	Conv4-64F	51.19 $\pm$ 0.76	67.65 $\pm$ 0.63	-	-	52.42 $\pm$ 0.76	63.76 $\pm$ 0.64
DN4 [43]	Conv4-64F	51.24 $\pm$ 0.74	71.02 $\pm$ 0.64	-	-	46.84 $\pm$ 0.81	74.92 $\pm$ 0.64
MeTAL [44]	Conv4-64F	52.63 $\pm$ 0.37	70.52 $\pm$ 0.29	-	-	-	-
PL [45]	Conv4-64F	48.00 $\pm$ 0.24	67.14 $\pm$ 0.23	-	-	60.11 $\pm$ 0.26	79.07 $\pm$ 0.25
LLDGP [46]	Conv4-64F	-	-	64.17 $\pm$ 0.31	78.42 $\pm$ 0.26	-	-
PSST [47]	Conv4-64F	-	-	64.37 $\pm$ 0.33	80.42 $\pm$ 0.32	-	-
HGNN [48]	Conv4-64F	55.63 $\pm$ 0.20	72.48 $\pm$ 0.16	-	-	69.02 $\pm$ 0.22	83.20 $\pm$ 0.15
ProtoNet+norm [15]	Conv4-64F	50.29 $\pm$ 0.41	67.13 $\pm$ 0.34	-	-	-	-
MC [16]	Conv4-64F	49.64 $\pm$ 0.83	65.67 $\pm$ 0.70	-	-	-	-
Transfer learning paradigms							
Baseline++ [17]	Conv4-64F	48.24 $\pm$ 0.75	66.43 $\pm$ 0.63	-	-	60.53 $\pm$ 0.83	79.34 $\pm$ 0.61
Neg-Cosine [24]	Conv4-64F	52.84 $\pm$ 0.76	70.41 $\pm$ 0.66	-	-	-	-
RFS-distill [29]	Conv4-64F	47.71	65.40	-	-	-	-
SKD [30]	Conv4-64F	48.14	66.36	-	-	-	-
CGCS [49]	Conv4-64F	55.53 $\pm$ 0.20	72.12 $\pm$ 0.16	-	-	-	-
Ours	Conv4-64F	57.99 $\pm$ 0.44	73.30 $\pm$ 0.36	67.02 $\pm$ 0.92	80.98 $\pm$ 0.69	69.92 $\pm$ 0.46	85.27 $\pm$ 0.28

- Indicates that the method described in the literature was not evaluated on certain datasets.

Table 4 indicates that (1) on MiniImageNet, CIFAR-FS, and CUB datasets, our method outperforms all SOTA methods based on lightweight models in both five-way one-shot and five-way five-shot classification tasks. (2) On MiniImageNet, among all the meta-learning-based SOTA methods, HGNN exhibits the best performance in few-shot image classification. Compared to HGNN, our method still achieves higher accuracy by 2.36% and 0.82% in the one-shot and five-shot tasks, respectively. Among the transfer learning-based SOTA methods, CGCS shows the best performance. In comparison to CGCS, our method also excels by 2.46% and 1.18% in the one-shot and five-shot tasks, respectively. (3) On CIFAR-FS, PSST stands out as the best performer. Our method surpasses PSST by 2.65% and 0.56% in the one-shot and five-shot tasks, respectively. (4) On CUB, our method achieves the best results among lightweight models in both one-shot and five-way tasks. It surpasses HGNN by 0.9% in the one-shot task and CGCS by 0.9% in the same task. These results demonstrate the effectiveness of our method in enhancing few-shot image classification capabilities of lightweight models on both generic and fine-grained classification datasets.

4.5. Detailed Analysis

4.5.1. Parameter Analysis

In Section 3.4, we detail the optimization of the student model's comprehensive loss function, outlined in Equation (22), with a particular focus on the coefficients α_1 and α_2 , which are crucial to the model's performance. We embarked on an extensive grid search to ascertain the optimal values for these coefficients, exploring a parameter range from 0.1 to 1.0 for both α_1 and α_2 . This methodical search was designed to not only individually

assess the impact of each coefficient but also to understand their combined effect on the classification performance through a joint analysis.

The results of this combined parameter search are visually summarized in Figure 4, where we employ heatmaps to illustrate the classification accuracy across the spectrum of α_1 and α_2 values. These heatmaps clearly highlight the regions of optimal performance, identified by the deepest red squares, representing the most effective parameter combinations for our model. Although there is a slight variation in the optimal values of α_1 and α_2 across different datasets, the patterns observed for one-shot and five-shot settings are consistent, reinforcing the reliability of our findings. Specifically, the optimal settings on MiniImageNet are identified as $\alpha_1 = 0.3$ and $\alpha_2 = 0.5$, whereas for CIFAR-FS and CUB, the best-performing values are both α_1 and α_2 at 0.5.

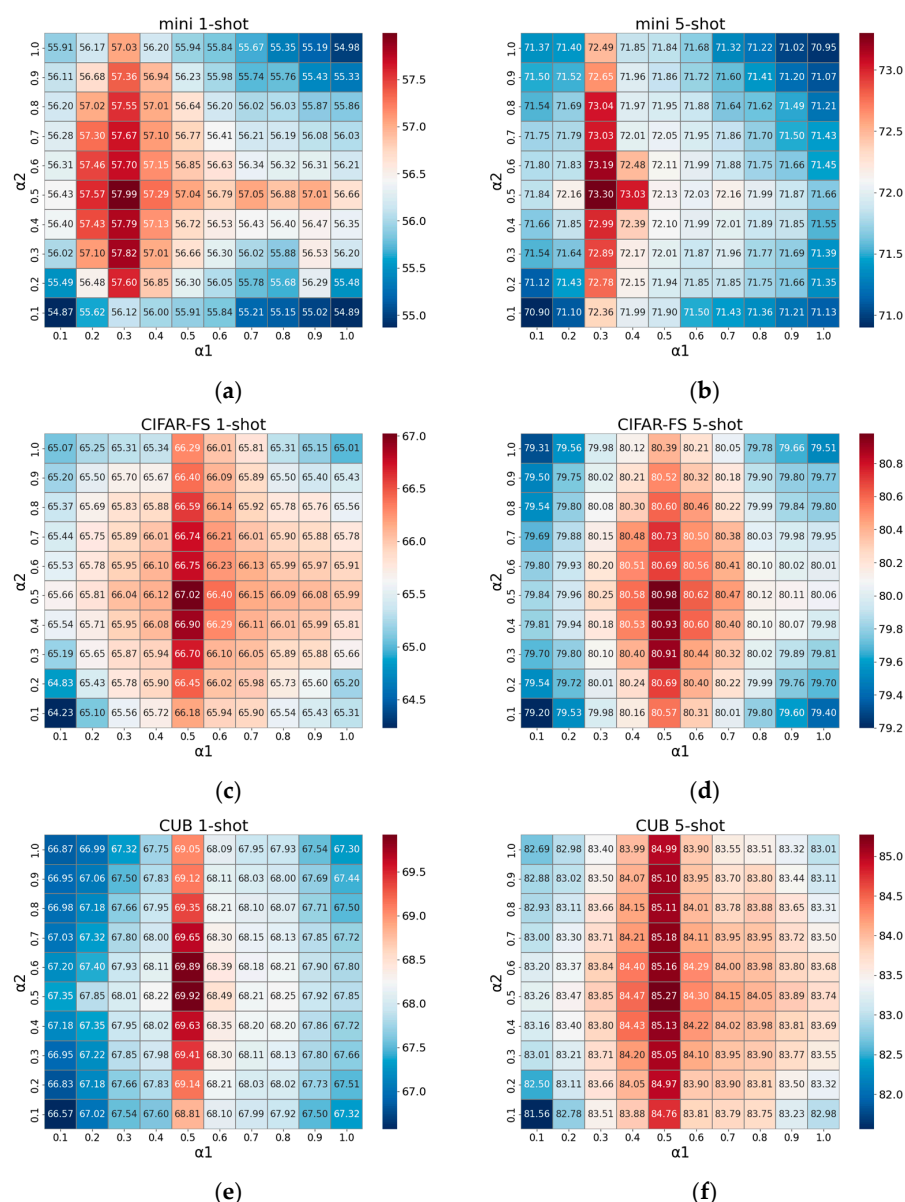


Figure 4. Heatmap visualization of α_1 and α_2 grid search on three datasets. (a) Grid search results on MiniImageNet in 1-shot task; (b) grid search results on MiniImageNet in 5-shot task; (c) grid search results on CIFAR-FS in 1-shot task; (d) grid search results on CIFAR-FS in 5-shot task; (e) grid search results on CUB in 1-shot task; (f) grid search results on CUB in 5-shot task.

Furthermore, we extended our exploration of α_1 and α_2 beyond the standard $[0, 1]$ range for the rigor of the experimental results. However, this additional analysis did not reveal any superior results, confirming that the optimal parameter values are indeed within the initial search range.

Additionally, Equation (16) from Section 3.3.2 introduces β , which modulates the skewness in Gaussian feature error calibration. Testing β values in the $[0, 2]$ range by 0.2 steps, we found, as demonstrated in Figure 5, that 0.8 delivers optimal performance across all datasets and tasks. Consequently, β was set to 0.8 for all further analyses.

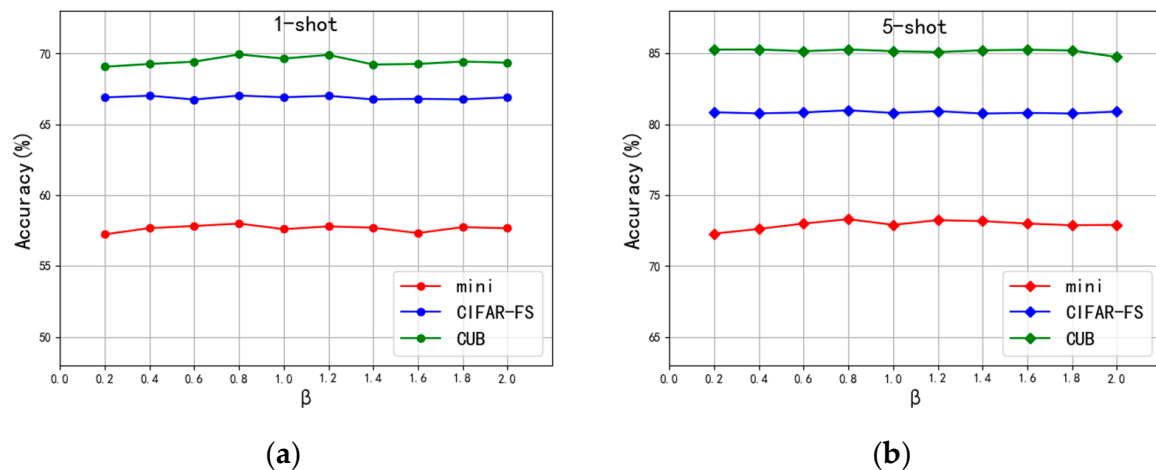


Figure 5. Test experiments under different values of β on three datasets: (a) 1-shot test accuracy under different values of β ; (b) 5-shot test accuracy under different values of β .

4.5.2. Visualization Analysis

This paper introduces a novel algorithm using a dual-faceted knowledge distillation strategy aimed at enhancing the feature extraction capabilities of lightweight student models by learning from a more complex teacher model. To evaluate our method's effectiveness, we applied the t-SNE [50] visualization technique to features extracted by the student model from test images. We selected five categories from the MiniImageNet and CIFAR-FS novel class datasets, randomly choosing 200 image samples per category. The t-SNE visualization results are presented in Figure 6, which includes both the baseline and our method.

The t-SNE visualizations illustrate that our method achieves a more distinct separation of feature points across different categories compared to the baseline, which relies solely on cross-entropy loss and pre-training on base class datasets. Specifically, the visualization indicates that our method produces feature points that are more tightly clustered within the same category and more widely spaced between different categories. For instance, the overlap of red and green points is notably less in our method compared to the baseline on MiniImageNet. On CIFAR-FS, the feature points pertaining to our method are significantly more clustered, with a greater separation between clusters of different colors, suggesting an enhanced feature distinction capability. This visual comparison underscores our method's enhanced capability in extracting distinctive features, thereby significantly advancing the performance in FSC tasks.

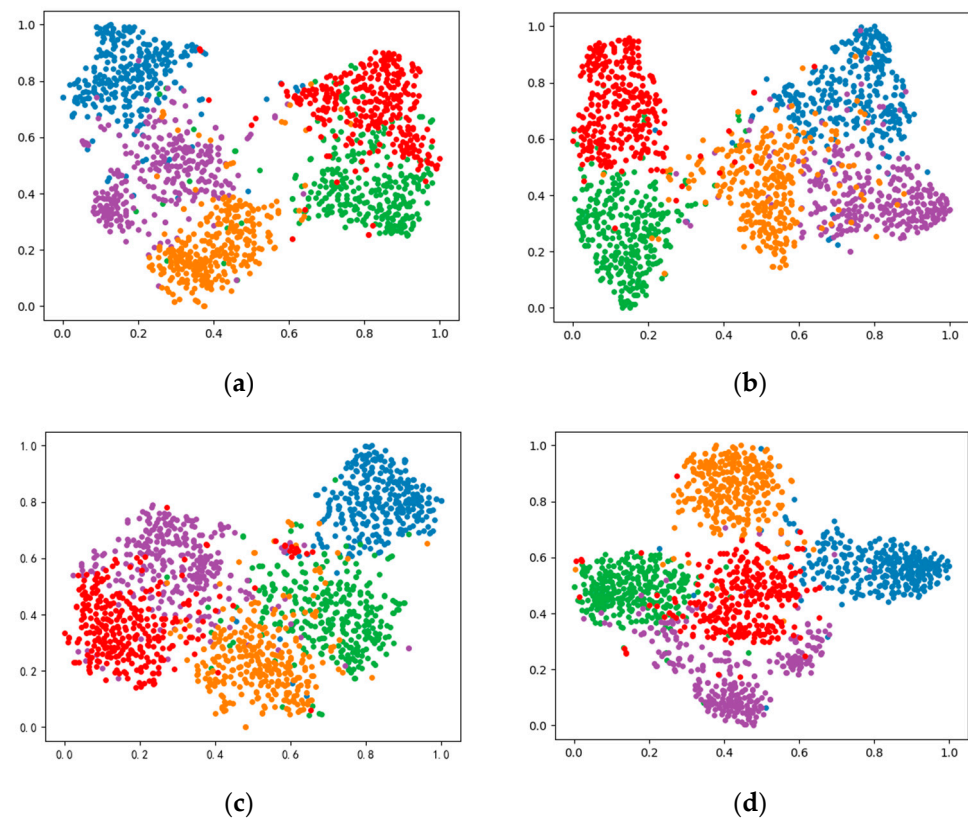


Figure 6. The t-SNE visualization of feature distributions on MiniImageNet and CIFAR-FS. (a) The feature distributions of 5 randomly selected novel classes from MiniImageNet by the baseline method; (b) the feature distributions of 5 randomly selected novel classes from MiniImageNet by our method; (c) the feature distributions of 5 randomly selected novel classes from CIFAR-FS by the baseline method; (d) the feature distributions of 5 randomly selected novel classes from CIFAR-FS by our method, where each color represents a distinct category of features.

5. Conclusions

This study presents a novel approach to enhancing few-shot image classification in lightweight models that is particularly suited for scenarios with limited storage and computational resources, such as mobile and embedded devices. Our key contribution is the introduction of a dual-faceted knowledge distillation strategy, effectively improving the performance of lightweight models. This strategy combines output-based and intermediate feature-based distillation techniques, augmented by a novel feature error distribution calibration method. This calibration addresses the limitations of traditional mean squared error functions in knowledge distillation, ensuring a closer approximation to the teacher model's feature distribution. Furthermore, we provide theoretical proof within the paper to substantiate this approach. This advancement not only enhances the understanding of knowledge distillation processes but also offers a robust framework for future research in this area.

Experiments conducted on multiple datasets, including MiniImageNet, demonstrate the superior performance of our method. Through rigorous ablation studies and comparisons with classical knowledge distillation algorithms, we have validated the enhanced base class classification and feature extraction capabilities of lightweight models using our approach. Notably, our method not only outperforms several established knowledge distillation algorithms but also surpasses current state-of-the-art methods based on lightweight models in few-shot image classification tasks.

In conclusion, the proposed model compression technique, leveraging a dual-faceted knowledge distillation strategy, offers a significant advancement in the field of few-shot

learning for lightweight models. This advancement has potential applications in various real-world scenarios where computational efficiency and model size are critical considerations.

Author Contributions: Conceptualization, B.Z. and J.G.; methodology, B.Z.; software, C.Y. and L.J.; validation, B.Z. and J.Z.; formal analysis, B.Z. and T.C.; investigation, X.Z. and T.C.; resources, J.Z.; data curation, X.Z.; writing—original draft preparation, B.Z.; writing—review and editing, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China under grant No. U2066203, No. 61973178, and No. 52377117, and in part by the Key Research & Development Program of Jiangsu Province under grant No. BE2021063.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in our study are publicly available datasets in the field of few-shot learning.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xie, J.; Long, F.; Lv, J.; Wang, Q.; Li, P. Joint distribution matters: Deep Brownian distance covariance for few-shot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
2. Bateni, P.; Goyal, R.; Masrani, V.; Wood, F.; Sigal, L. Improved few-shot visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
3. Bogdan, T.N.; Bruno, M.; Rafael, W.; Victor, B.G.; Vanderlei, Z.; Lourival, L. A computer vision system for monitoring disconnect switches in distribution substations. *IEEE Trans. Power Deliv.* **2022**, *37*, 833–841.
4. Liu, Y.; Zhang, W.; Xiang, C.; Zheng, T.; Cai, D.; He, X. Learning to affiliate: Mutual centralized learning for few-shot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
5. Yang, Z.; Wang, J.; Zhu, Y. Few-shot classification with contrastive learning. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
6. Lai, J.; Yang, S.; Zhou, J.; Wu, W.; Chen, X.; Liu, J.; Gao, B.; Wang, C. Clustered-patch element connection for few-shot learning. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 19–25 August 2023.
7. Ma, J.; Zhang, Y.; Ma, Z.; Mao, K. Research progress of lightweight neural network convolution design. *J. Front. Comput. Sci. Technol.* **2022**, *16*, 512–528.
8. Song, Y.; Wang, T.; Cai, P.; Mondal, S.K.; Sahoo, J.P. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* **2023**, *55*, 1–40. [[CrossRef](#)]
9. Zhang, M.; Zhang, J.; Lu, Z.; Xiang, T.; Ding, M.; Huang, S. IEPT: Instance-level and episode-level pretext tasks for few-shot learning. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 3–7 May 2021.
10. Li, W.; Wang, Z.; Yang, X.; Dong, C.; Tian, P.; Qin, T.; Huo, J.; Shi, Y.; Wang, L.; Gao, Y.; et al. LibFewShot: A comprehensive library for few-shot learning. *Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 14938–14955. [[CrossRef](#)] [[PubMed](#)]
11. Liu, Y.; Lei, Y.; Fan, J.; Wang, F.; Gong, Y.; Tian, Q. Survey on image classification technology based on small sample learning. *Acta Autom. Sin.* **2021**, *42*, 297–315.
12. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep network. In Proceedings of the International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017.
13. Mike, H.; Jan, N.R.; Aske, P. A survey of deep meta-learning. *Artif. Intell. Rev.* **2021**, *54*, 4483–4541.
14. Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; Cord, M. Boosting few-shot visual learning with self-supervision. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
15. Bouniot, Q.; Redko, I.; Audigier, R.; Loesch, A.; Habrard, A. Kernel relative-prototype spectral filtering for few-shot learning. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
16. Ma, R.; Fang, P.; Avraham, G.; Zuo, Y.; Zhu, T.; Drummond, T.; Harandi, M. Learning instance and task-aware dynamic kernels for few-shot learning. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
17. Chen, W.; Liu, Y.; Kira, Z.; Wang, Y.F.; Huang, J. A closer look at few-shot classification. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
18. Liu, B.; Cao, Y.; Lin, Y.; Zhang, Z.; Long, M.; Hu, H. Negative margin matters: Understanding margin in few-shot classification. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, Scotland, 23–23 August 2020.

19. Mangla, P.; Singh, M.; Sinha, A.; Kumari, N.; Balasubramanian, V.; Krishnamurthy, B. Charting the right manifold: Manifold mixup for few-shot learning. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2020.
20. Rizve, M.N.; Khan, S.; Khan, F.S.; Shah, M. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Conference, 19–25 June 2021.
21. Xu, J.; Pan, X.; Luo, X.; Pei, W.; Xu, Z. Exploring category-correlated feature for few-shot image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
22. Geoffrey, H.; Oriol, V.; Jeff, D. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
23. Adriana, R.; Nicolas, B.; Samira, E.K.; Antoine, C.; Carlo, G.; Yoshua, B. Fitnets: Hints for thin deep nets. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 5–7 May 2015.
24. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
25. Tung, F.; Mori, G. Similarity-preserving knowledge distillation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
26. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
27. Passalis, N.; Tzelepi, M.; Tefas, A. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2030–2039. [[CrossRef](#)] [[PubMed](#)]
28. Huang, Z.; Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv* **2017**, arXiv:1707.01219.
29. Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J.B.; Isola, P. Rethinking few-shot image classification: A good embedding is all you need? In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, Scotland, 23–23 August 2020.
30. Rajasegaran, J.; Khan, S.; Hayat, M.; Khan, F.S.; Shah, M. Self-supervised knowledge distillation for few-shot learning. *arXiv* **2020**, arXiv:2006.09785.
31. Ma, J.; Xie, H.; Han, G.; Chang, S.F.; Galstyan, A.; Abd-Almageed, W. Partner-assisted learning for few-shot image classification. In Proceedings of the International Conference on Computer Vision (ICCV), Virtual Conference, 10–17 October 2021.
32. Zhou, Z.; Qiu, X.; Xie, J.; Wu, J.; Zhang, C. Binocular mutual learning for improving few-shot classification. In Proceedings of the International Conference on Computer Vision (ICCV), Virtual Conference, 10–17 October 2021.
33. Gomes, J.C.; Borges, L.d.A.B.; Borges, D.L. A multi-layer feature fusion method for few-Shot image classification. *Sensors* **2023**, *23*, 6880. [[CrossRef](#)] [[PubMed](#)]
34. Zhang, P.; Li, Y.; Wang, D.; Wang, J. RS-SSKD: Self-supervision equipped with knowledge distillation for few-shot remote sensing scene classification. *Sensors* **2021**, *21*, 1566. [[CrossRef](#)] [[PubMed](#)]
35. Tukey, J.W. *Exploratory Data Analysis*; Addison-Wesley: Reading, MA, USA, 1977.
36. Li, K.; Zhang, Y.; Li, K.; Fu, Y. Adversarial feature hallucination networks for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
37. Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K. Meta-learning for semi-supervised few-shot classification. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
38. Hilliard, N.; Phillips, L.; Howland, S.; Yankov, A.; Corley, C.D.; Hodas, N.O. Few-shot learning with metric-agnostic conditional embeddings. *arXiv* **2018**, arXiv:1802.04376.
39. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
40. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018.
41. Bertinetto, L.; Henriques, J.F.; Torr, P.; Vedaldi, A. Meta-learning with differentiable closed-form solvers. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
42. Li, W.; Xu, J.; Huo, J.; Wang, L. Distribution consistency based covariance metric networks for few-shot learning. In Proceedings of the American Association for Artificial Intelligence (AAAI), Hawaii, HI, USA, 27 January–1 February 2019.
43. Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; Luo, J. Revisiting local descriptor-based image-to-class measure for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
44. Baik, S.; Choi, J.; Kim, H.; Cho, D.; Min, J.; Lee, K.M. MetaLearning with task-adaptive loss function for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Conference, 19–25 June 2021.
45. Snell, J.; Zemel, R. Bayesian few-shot classification with one-vs-each poly-gamma augmented gaussian processes. *arXiv* **2020**, arXiv:2007.10417.
46. Wang, Z.; Miao, Z.; Zhen, X.; Qiu, Q. Learning to learn dense gaussian processes for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual Conference, 6–14 December 2021.
47. Chen, Z.Y.; Ge, J.X.; Zhan, H.S.; Huang, S.; Wang, D.L. Pareto self-supervised training for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Conference, 19–25 June 2021.

48. Yu, T.; He, S.; Song, Y.Z.; Xiang, T. Hybrid graph neural networks for few-shot learning. In Proceedings of the American Association for Artificial Intelligence (AAAI), Vancouver, BC, Canada, 22 February–1 March 2022.
49. Gao, Z.; Wu, Y.; Jia, Y.; Harandi, M. Curvature generation in curved spaces for few-shot learning. In Proceedings of the International Conference on Computer Vision (ICCV), Virtual Conference, 10–17 October 2021.
50. Laurens, L.M.; Geoffrey, H. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.