Boosting Few-Shot Learning via Attentive Feature Regularization

Xingyu Zhu[1, 2], **Shuo Wang**[1, 2*], **Jinda Lu**[1, 2], **Yanbin Hao**[1, 2], **Haifeng Liu**[3], **Xiangnan He**[1, 2]

[1]Department of Electronic Engineering and Information Science, University of Science and Technology of China;
[2]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China;
[3]Brain-Inspired Technology Co., Ltd.
{xingyuzhu, lujd}@mail.ustc.edu.cn, {shuowang.edu, xiangnanhe}@gmail.com,
haoyanbin@hotmail.com, haifeng@leinao.ai

Abstract

Few-shot learning (FSL) based on manifold regularization aims to improve the recognition capacity of novel objects with limited training samples by mixing two samples from different categories with a blending factor. However, this mixing operation weakens the feature representation due to the linear interpolation and the overlooking of the importance of specific channels. To solve these issues, this paper proposes attentive feature regularization (AFR) which aims to improve the feature representativeness and discriminability. In our approach, we first calculate the relations between different categories of semantic labels to pick out the related features used for regularization. Then, we design two attention-based calculations at both the instance and channel levels. These calculations enable the regularization procedure to focus on two crucial aspects: the feature complementarity through adaptive interpolation in related categories and the emphasis on specific feature channels. Finally, we combine these regularization strategies to significantly improve the classifier performance. Empirical studies on several popular FSL benchmarks demonstrate the effectiveness of AFR, which improves the recognition accuracy of novel categories without the need to retrain any feature extractor, especially in the 1-shot setting. Furthermore, the proposed AFR can seamlessly integrate into other FSL methods to improve classification performance.

Introduction

In recent years, convolutional neural networks (CNNs) have demonstrated remarkable capabilities on various visual classification tasks, particularly provided with sufficient training data. However, collecting and labeling such datasets is a time-consuming and expensive procedure. As a remedy to address this challenge, few-shot learning (FSL) is proposed to classify a novel object with a scarcity of labeled data. (Ye et al. 2020; Peng et al. 2019; Wang et al. 2020).

The conventional solution of FSL involves using a CNN trained on the base categories to directly extract the global features of novel objects (Hariharan and Girshick 2017; Wang et al. 2018). It aims to yield a transferable feature representation (textures and structures) to describe a novel category. Subsequently, these features are employed to train a classifier for recognizing novel objects. Manifold regularization (Rodríguez et al. 2020; Deutsch et al. 2017; Velazquez et al. 2022) is a popular strategy to improve classification performance. These methods involve mixing two samples (features) and their labels from randomly selected categories to generate a regularized feature. However, the mixing operation with randomness is easy to weaken the representation ability (Guo, Mao, and Zhang 2019; Chou et al. 2020). This is primarily due to the direct interpolation without considering the complementarity of two features and the neglect of specific feature channels (Hou, Liu, and Wang 2017; Shi, Wu, and Wang 2023; Luo, Xu, and Xu 2022; Zhu et al. 2023), which in turn impacts the distribution of prediction results. As illustrated in Figure 1, given a novel sample of "Retriever" and another randomly picked out sample "Linnet", the manifold regularization methods, *e.g.*, Mixup (Zhang et al. 2018), CutMix (Yun et al. 2019), and PatchMix (Liu et al. 2021), interpolate their images and labels to train the classifier for predicting both categories. It's evident that the "Retriever" and the "Linnet" are unrelated in terms of both vision and semantics. Consequently, the regularized features deviate from the novel feature "Retriever" (as indicated by the five yellow squares in the lower-left corner of Figure 1). This deviation leads to an increase in the prediction score for "Linnet" and results in misclassification. This deviation leads to an increase in the prediction score of the "Linnet" and limits the classification results.

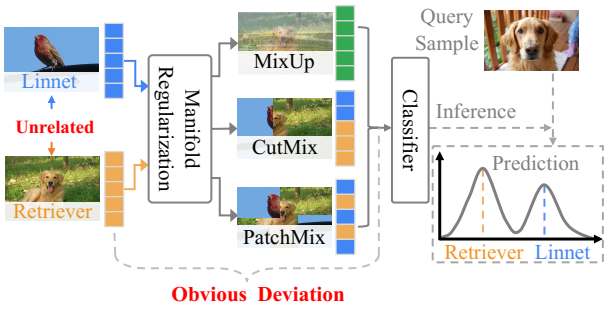

Figure 1: The analysis of manifold regularization methods.

To address the aforementioned issue arising from manifold regularization, we first incorporate semantics to select categories related to the novel categories from the base set.

*Corresponding author

This idea aligns with the prior work (Wang et al. 2020; Peng et al. 2019; Wang et al. 2022), where semantic knowledge not only strengthens visual features but also aids help classifier in capturing discriminative patterns. However, it's worth noting that such methodologies necessitate greater prior semantic information during the training process, which leads to increased model size and longer training times. Different from the previous approaches, our method solely relies on semantic labels to select relevant base categories during the data preprocessing stage. This purposeful selection, in contrast to the random selection in manifold regularization, enables the classifier to better concentrate more effectively on the novel content during the training stage. Besides, we also exploit the feature complementarity from similar categories and the discriminability of specific feature channels, which can both provide distinctive patterns for classification (Liu et al. 2019; Shi, Wu, and Wang 2023). Building on the above analysis, we propose two attention-based calculations at the instance and channel levels, respectively.

The instance attention is designed to adaptively leverage the collaborative components of the selected base categories guided by their relevance to heighten the novel feature representativeness. Specifically, we first analyze the semantic similarity of the selected base categories related to the novel category, then calculate the attention scores between the selected base samples and the given novel sample to measure their importance. Finally, these attention scores are then employed in the reweighting of selected samples. Instance attention exploits the collaboration between the related categories through adaptive interpolation, which avoids the irrelevant components of the base categories and consequently improves the representation of the novel samples.

For channel calculations, we aim to emphasize the specific feature channels that signify the discriminative patterns. Specifically, we calculate the scores as channel importance weights from the regularized features output from the instance attention. These weights are then applied to each channel of features in regularization, aiding the classifier in identifying the representative content of the novel samples. This channel attention mechanism allows for more efficient and focused exploration of novel category information within the feature channels, which enhances the discriminability of the final feature representation.

The proposed procedures, defined as Attentive Feature Regularization (AFR), all operate on features and can be easily applied to existing pre-trained feature extractors. The main contributions of our method are as follows.

1. We propose instance-level attention with semantic selection to improve the feature representativeness, which leverages the complementarity of the related base categories to enhance the novel categories.
2. We design channel-level attention to enhance the feature discriminability by measuring the importance of different channels, which helps the classifier focus on the representative content of the novel sample.
3. Our method achieves state-of-the-art performance on three popular FSL datasets and can also be used to improve the performance of the classifier in other FSL methods without training feature extractors.

Related Work

In this section, we first briefly introduce common solutions for FSL tasks and corresponding regularization strategies. Subsequently, we list the applications of recent attention-based methods. Finally, we enumerate the differences between our methods and those of related methods.

Knowledge Transfer in Few-Shot Learning

Recent advances in Few-Shot Learning (FSL) have demonstrated promising performance by transferring the knowledge from the base categories to the novel categories (Li et al. 2020, 2019; Wang et al. 2022; Lu et al. 2023). These methods leverage semantic knowledge to provide additional information for refining visual features or enriching the supervision during classifier training. For example, the method in (Li et al. 2019) clusters hierarchical textual labels from both the base and novel categories to improve the feature extractor training. Wang *et al.* proposed a multi-directional knowledge transfer (MDKT) method which integrates the visual and textual features through a bidirectional knowledge connection. The work described in (Lu et al. 2023) employs the semantics to explore the correlation of categories to hallucinate the additional training samples.

Regularization in Few-Shot Learning

Recently, manifold regularization(Devries and Taylor 2017; Zhang et al. 2018; Verma et al. 2019; Yun et al. 2019; Liu et al. 2021) has been used in FSL tasks, which is simply based on mixture and mask operation and can improve the classification performance. The simplest method is CutOut (Devries and Taylor 2017), which randomly masks out square regions of input during training and improves the performance of the networks. Based on CutOut, many other manifold regularization methods have been developed, *i.e.*, MixUp (Zhang et al. 2018; Verma et al. 2019), CutMix (Yun et al. 2019), PatchMix (Liu et al. 2021). Specifically, MixUp mixes two samples by interpolating both the image and the labels. In CutMix, patches are cut and pasted among training features, where ground truth labels are also mixed proportionally to the area of the patches. PatchMix is similar to CutMix and uses mixed images for contrastive learning.

Attention in Few-Shot Learning

In the field of FSL, attention mechanisms (Vaswani et al. 2017a) have been widely widespread due to their ability to highlight the important parts of inputs by measuring similarities. This enables the network to focus on critical content for specific tasks (Hou et al. 2019; Kang et al. 2021; Chikontwe, Kim, and Park 2022). For instance, Hou *et al.* proposed a cross-attention (CAM) method to model the semantic relevance between class and query features, leading to adaptive localization of relevant regions and generation of more discriminative features (Hou et al. 2019). The work in (Kang et al. 2021) computes the cross-correlation between two representations and learns to produce co-attention between them. It improves the classification accuracy by learning

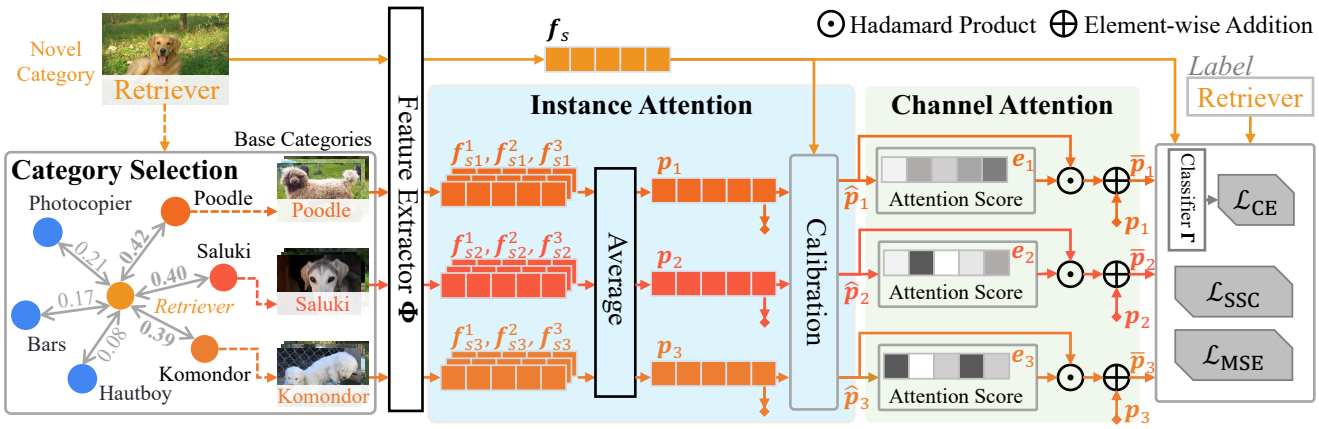

Figure 2: The overview of attentive feature regularization (AFR), where $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{SC}}$, and $\mathcal{L}_{\text{MSE}}$ are three losses.

cross-correlational patterns and adapting "where to attend" concerning the images given in the testing stage. Moreover, the method (Ye et al. 2020) integrates the entire Transformer module including attention mechanisms (FEAT) to adapt the features for FSL tasks. The authors in (Lai et al. 2022) propose the method named transformer-based Semantic Filter (tSF) which defines the additional learnable parameters to filter the useful knowledge of the whole base set for the novel category. The recently proposed CAD (Chikontwe, Kim, and Park 2022) employs self-attention operations to cross-attend support and query embeddings, effectively reweighting each instance relative to the others.

Based on the analysis of the related work, our method belongs to the manifold regularization methods. The methods most related to ours are the recently proposed Manifold Mixup in (Verma et al. 2019) and CAM in (Hou et al. 2019). Our method differs from theirs in two aspects. First, we introduce semantic knowledge to purposefully select samples for regularization and keep the label of the regularized feature the same as the novel feature to avoid introducing other unrelated supervisions for training. Second, we design two attention calculations to enhance collaboration and improve the discriminability of features, which helps the classifier focus on the distribution of novel categories rather than associate the support and query samples during the testing stage (Hou et al. 2019). Besides, our approach directly applies to the features and has a lower computational complexity.

Method

In this section, we elaborate on our attentive feature regularization (AFR). First, we briefly revisit the preliminaries of the FSL tasks and an overview of our framework. Second, we delve into the details of our semantic selection process and different attention calculations. Finally, we describe the training and inference procedures of our approach.

Preliminaries

The data for the few-shot learning task is divided into three parts: base set $\mathcal{D}_{\text{base}}$, support set $\mathcal{D}_{\text{support}}$, and query set $\mathcal{D}_{\text{query}}$. The base set $\mathcal{D}_{\text{base}}$ has large-scale labeled samples (*e.g.*, about hundreds of samples in one category) used for training the feature extractor. The categories of these samples are denoted as $\mathcal{C}_{\text{base}}$ and provide valuable prior knowledge as known contents to describe other samples. The support set $\mathcal{D}_{\text{support}}$ and the query set $\mathcal{D}_{\text{query}}$ share the same set of categories, called $\mathcal{C}_{\text{novel}}$, which is disjoint with that of the base set $\mathcal{C}_{\text{base}}$. The goal of few-shot learning is to construct a classifier using training samples from both the base set and the support set, capable of accurately classifying the samples in the query set. For the training samples from $\mathcal{D}_{\text{support}}$, there are total N categories that are randomly sampled from $\mathcal{C}_{\text{novel}}$, and each category provides K samples. This process is known as the N-way-K-shot recognition problem.

The overview of our framework is depicted in Figure 2. First, we use the semantic knowledge to select the related base categories to a given novel sample and extract features of all these samples by a pre-trained CNN. Second, we design instance attention and channel attention to regularize these features. Third, we design three losses to constrain the regularization procedure and train a classifier.

Attentive Feature Regularization

Textual knowledge uses semantic description to express each category. It provides the direct relations between the categories. To avoid bringing irrelevant noise to influence the classifier training, we directly calculate the relations between these descriptions before regularization. Specifically, we first use the word2vec embedding method (Li et al. 2019) to express these descriptions into the feature. Then, given feature of a support category as $\boldsymbol{t}_s$, we calculate the relations $\mathcal{R}^s = \{r_i^s\}_{i=1}^{|\mathcal{C}_{\text{base}}|}$ between $\boldsymbol{t}_s$ and the other descriptions $\{\boldsymbol{t}_i\}_{i\in\mathcal{C}_{\text{base}}}$ from base categories by similarity calculation:

$$r_i^s = \frac{\langle \boldsymbol{t}_s, \boldsymbol{t}_i \rangle}{\|\boldsymbol{t}_s\|_2 \cdot \|\boldsymbol{t}_i\|_2}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the inner product between two vectors.

After obtaining the relation scores $\mathcal{R}^s$, we sort them and select the samples from the top-β_s related categories denoted as $\mathcal{C}_{\beta_s}$ for regularization. These semantically relevant features can provide a more relevant content supplement to the training and avoid bringing much irrelevant noise.

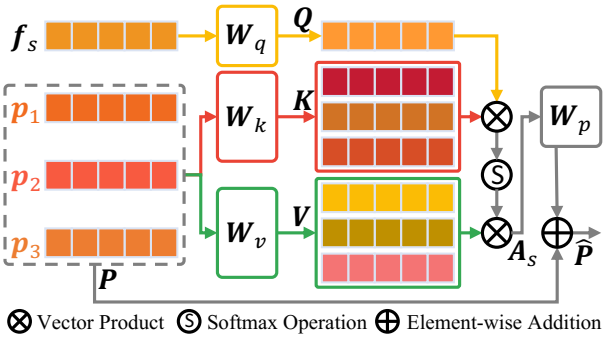

Figure 3: The calculation of calibration in instance attention.

Our regularization operates on the feature level. Therefore, we first represent the sample $\boldsymbol{I}$ into feature $\boldsymbol{f} = \boldsymbol{\Phi}(\boldsymbol{I}) \in \mathbb{R}^d$ by extracting the output from the pre-trained visual model $\boldsymbol{\Phi}$ before the last prediction layer. The model $\boldsymbol{\Phi}$ is already trained on known images from the base set $\mathcal{D}$base, and d is the dimension of the feature. Then, we illustrate the different attention calculations used in our approach.

Instance Attention Given a support feature $\boldsymbol{f}_s$ with its textual description feature $\boldsymbol{t}_s$, we first selected β_s categories from base categories as $\mathcal{C}_{\beta_s}$ using Eq (1). Then we compute the prototype of each selected category $\mathcal{C}^i_{\beta_s}$ by averaging all the features corresponding to that category:

$$\boldsymbol{p}_i = \frac{1}{|\mathcal{C}^i_{\beta_s}|} \sum \boldsymbol{f}_j,\ j \in \mathcal{C}^i_{\beta_s} \tag{2}$$

Therefore, the prototypes of whole $\mathcal{C}_{\beta_s}$ categories can be constructed as $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_{|\mathcal{C}_{\beta_s}|}]$, where $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{C}_{\beta_s}| \times d}$. We then design a calibration based on self-attention (Vaswani et al. 2017b) to find relevant categories that can help describe the novel category. The details are shown in Figure 3. Specifically, we first design three attention matrixes $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ to capture the content similarities from novel features and prototypes of related categories:

$$\boldsymbol{Q} = \boldsymbol{f}_s * \boldsymbol{W}_q, \boldsymbol{K} = \boldsymbol{P} * \boldsymbol{W}_k, \boldsymbol{V} = \boldsymbol{P} * \boldsymbol{W}_v, \tag{3}$$

where $\boldsymbol{W}_q \in \mathbb{R}^{d \times d}$, $\boldsymbol{W}_k \in \mathbb{R}^{d \times d}$, $\boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ are weights of calibration calculation. We then use these similarities to calibrate the prototypes of the related categories since the distribution of the base categories and the novel category belong to different spaces, where the amplitude $\boldsymbol{A}_s \in \mathbb{R}^d$ of calibration is calculated as:

$$\boldsymbol{A}_s = \text{softmax}(\frac{\boldsymbol{Q} * \boldsymbol{K}^\intercal}{\sqrt{d}})\boldsymbol{V}. \tag{4}$$

It measures the relations between the novel category and its related base categories from the feature space.

Thus, calibrated prototypes $\hat{\boldsymbol{P}}$ can be defined as:

$$\begin{aligned} \hat{\boldsymbol{P}} &= [\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2, ..., \hat{\boldsymbol{p}}_{|\mathcal{C}_{\beta_s}|}], \\ &= \delta(\boldsymbol{A}_s * \boldsymbol{W}_p) + [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_{|\mathcal{C}_{\beta_s}|}], \end{aligned} \tag{5}$$

where $\boldsymbol{W}_p \in \mathbb{R}^{d \times d}$ is weight matrix for the calibration calculation, and δ is ReLU function. The calibrated prototypes can better simulate the distribution of the novel category and improve the accuracy of feature regularization.

Channel Attention Channels of features have different influences on classifiers (Yue et al. 2020). To identify the important content of the channels, we design a channel attention module inspired by SE-Net (Hu, Shen, and Sun 2018). SE-Net terms the "Squeeze-and-Excitation (SE)" block to adaptively re-calibrate channel-wise feature responses by explicitly modeling interdependencies between channels in backbone training(Hu, Shen, and Sun 2018). Thus, we introduce a similar operation into the feature analysis. Specifically, we design two fully connected (FC) layers to "Squeeze-and-Excitation" calibrated prototypes $\hat{\boldsymbol{P}}$:

$$\boldsymbol{E_s} = \sigma(\boldsymbol{FC}_2(\delta(\boldsymbol{FC}_1(\hat{\boldsymbol{P}})))), \tag{6}$$

where σ is the Sigmoid function, and we intentionally set the embedding size of $\boldsymbol{FC}_1$ to be smaller than that of $\boldsymbol{FC}_2$. This design enhances important content and weakens unrelated content in the features by controlling the size of $\boldsymbol{FC}_1$. To further fuse the channel attention with the prototypes, we set the size of $\boldsymbol{E_s}$ to be the same as $\hat{\boldsymbol{P}} \in \mathbb{R}^{|\mathcal{C}_{\beta_s}| \times d}$ by controlling $\boldsymbol{FC}_2$ accordingly. In our fusion stage, we employ a residual structure to prevent vanishing gradients while improving the accuracy of the prototype representation:

$$\bar{\boldsymbol{P}} = \boldsymbol{E_s} \odot \hat{\boldsymbol{P}} + \boldsymbol{P}, \tag{7}$$

where $\odot$ is the Hadamard product. $\bar{\boldsymbol{P}} \in \mathbb{R}^{|\mathcal{C}_{\beta_s}| \times d}$ not only close to the distribution of the novel category by calibrating but also captures the content related to the novel category by using channel attention. Therefore, we sample the features of $\bar{\boldsymbol{P}}$ as representations of the given novel category to enrich the training set in our few-shot learning task.

Training and Inference

Denoted the novel samples and their labels in a N-way-K-shot task as $\{\{\boldsymbol{f}^i_s, \boldsymbol{l}^i_s\}^N_{s=1}\}^K_{i=1}$, and the fused prototypes as $\{\bar{\boldsymbol{P}}_s = \{\bar{\boldsymbol{p}}^j_s\}^{\beta_s}_{j=1}\}^N_{s=1}$, we combine the given features and prototypes into one set $\{\{\boldsymbol{H}_s = \{\boldsymbol{h}^j_s\}\}^N_{s=1}\}^{K+\beta_s}_{j=1}$ to simplify the expressions in subsequent calculations, where $\boldsymbol{H}_s = [\boldsymbol{f}^1_s, \boldsymbol{f}^2_s, ..., \boldsymbol{f}^K_s, \bar{\boldsymbol{p}}^1_s, \bar{\boldsymbol{p}}^2_s, ..., \bar{\boldsymbol{p}}^{\beta_s}_s]$. We then design two losses to constrain the distribution of regularized prototypes and use cross-entropy (CE) loss to train the classifier.

First, we adopt the principles of self-supervised contrastive learning (Khosla et al. 2020), which aim to bring features of the same category closer together while pulling features of different categories apart. Thus, the supervised contrastive (SC) loss can be calculated as follows:

$$\mathcal{L}_{\text{SC}} = \frac{1}{N|\boldsymbol{H}_s|} \sum_{s=1}^{N} \sum_{\substack{i,j=1 \\ i \neq j}}^{|\boldsymbol{H}_s|} \log \frac{\exp(\langle \boldsymbol{h}^i_s, \boldsymbol{h}^j_s \rangle / \tau)}{\sum_{\forall \boldsymbol{h}_p \notin \boldsymbol{H}_s} \exp(\langle \boldsymbol{h}^i_s, \boldsymbol{h}_p \rangle / \tau)}, \tag{8}$$

where $|\boldsymbol{H}_s| = K + \beta_s$ means the size of $\boldsymbol{H}_s$. Minimizing $\mathcal{L}_{\text{SC}}$ encourages maximizing the distances between samples from different categories and clustering them from the same category closer together. Meanwhile, to bridge the distribution gap between the prototypes of base categories and the features of the novel category, we employ the mean squared

error (MSE) operation to measure the average prototypes and the novel features, and the loss is designed as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{s=1}^{N}||\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{f}_s^i - \frac{1}{\beta_s}\sum_{j=1}^{\beta_s}\bar{\boldsymbol{p}}_s^j||. \tag{9}$$

Finally, we design a N-way classifier $\boldsymbol{\Gamma}$ to learn the prediction distribution from given novel features and the prototypes. In this work, the classifier $\boldsymbol{\Gamma}$ is a simple network, *e.g.*, as simple as one fully connected layer. We use the cross-entropy loss to train the classifier with the hard labels:

$$\mathcal{L}_{\text{CE}} = \frac{1}{N}\frac{1}{|\boldsymbol{H}_s|}\sum_{s=1}^{N}\sum_{i=1}^{|\boldsymbol{H}_s|}\text{CrossEntropy}(\boldsymbol{h}_s^i, \boldsymbol{l}_s). \tag{10}$$

The total loss for training is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mu_1\mathcal{L}_{\text{SC}} + \mu_2\mathcal{L}_{\text{MSE}}, \tag{11}$$

where μ_1 and μ_2 are two weighting factors.

During the inference, we use the trained classifier to directly predict the category for each feature in the query set.

Experiments

In this section, we present the experimental evaluation of our AFR. We begin by introducing the experimental settings. Next, we perform ablation studies to analyze the contributions of different components in our approach. Finally, we compare the performance of our approach with other state-of-the-art (SOTA) methods. Our experiments aim to address the following research questions (**RQ**):

RQ1: Given an novel category, how many related categories ($\mathcal{C}_{\beta_s}$) should be selected from the base categories?

RQ2: What are the effects of the instance attention and channel attention?

RQ3: How do the contrastive learning and the feature space closing operations influence the classifier?

RQ4: How does AFR perform compared to the state-of-the-art FSL methods?

Experimental Settings

Datasets. We evaluate our method on three benchmark datasets, *i.e.*, Mini-ImageNet (Vinyals et al. 2016), Tiered-ImageNet (Ren et al. 2018), and Meta-Datase (Triantafillou et al. 2019). Specifically, Mini-ImageNet consists of 100 categories and each category has 600 images. It is divided into three parts: 64 base categories for training, 16 novel categories for validation, and the remaining 20 categories for testing. Similar to Mini-ImageNet, Tiered-ImageNet consists of 779165 images from 608 categories, where 351 base categories are used for training, 97 novel categories are used for validation, and the remaining 160 novel categories are used for testing. Meta-Dataset is a significantly larger-scale dataset that comprises multiple datasets with diverse data distributions, and we follow the usage described in (Xu et al. 2022). Specifically, feature extractor training is conducted using the base categories of Mini-ImageNet, and the other 8 image datasets are utilized for testing process, including Omniglot (Lake, Salakhutdinov, and Tenenbaum 2015), CUB-200-2011 (Wah et al. 2011), Describable Textures (Cimpoi et al. 2014), Quick Draw (Fernandez-Fernandez et al. 2019), Fungi (Sulc et al. 2020), VGG Flower (Nilsback and Zisserman 2008), Traffic Signs (Houben et al. 2013).

Evaluation. In our evaluation, we conduct several N-way-K-shot classification tasks. In each task, N novel categories are randomly sampled at first, then K samples in each of the N categories are sampled for training, and finally, 15 samples (different from the previous K samples) in each of the N categories are sampled for testing. To ensure reliable results, we sample 600 such tasks and report mean accuracies and variances on all tasks. In our experiments, $N = 5$. Notably, we adhere to the evaluation setting for meta-dataset as described in (Xu et al. 2022), where the novel categories are randomly sampled from the alternate image datasets, excluding the base categories present in the Mini-ImageNet.

Implementation Details. We utilize the features extracted from the pre-trained model and then apply our AFR to obtain both original and regularized features for training the classifier $\boldsymbol{\Gamma}$. These features are used to train the classifier $\boldsymbol{\Gamma}$ using the loss function $\mathcal{L}$ defined in Eq. (11) for a total of 1000 epochs. We employ Adam optimization (Kingma and Ba 2015) with a learning rate of 0.001 and a weight decay of 0.0001 during the training process.

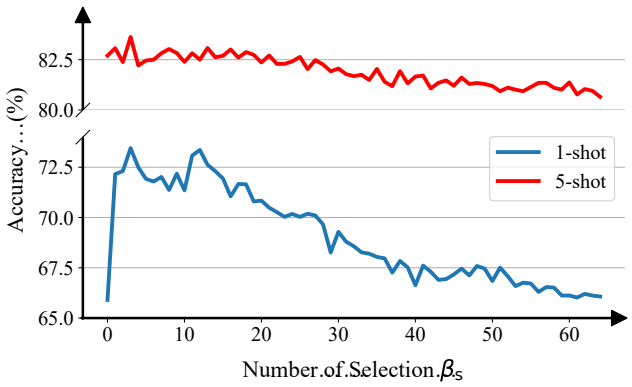

Figure 4: The accuracy (%) of the classifiers trained with the different numbers of selected base categories.

Ins.Att.	*Chanl.Att*	$K = 1$	$K = 5$
✗	✗	64.02 ± 0.70%	82.32 ± 0.41%
✗	✓	68.74 ± 0.61%	82.56 ± 0.46%
✓	✗	68.03 ± 0.58%	83.04 ± 0.43%
✓	✓	**70.68** ± 0.61%	**83.36** ± 0.45%

Table 1: The accuracy (%) of the classifiers with different attentions, where *Ins.Att.* and *Chanl.Att* is the instance attention and channel attention, respectively.

Ablation Study

In the ablation study, we use 64 base categories and 16 novel categories (validation set) of Mini-ImageNet with the available ResNet-12 (Chen et al. 2021) to evaluate the effective-

$\mathcal{L}_{\mathrm{SC}}$	$\mathcal{L}_{\mathrm{MSE}}$	$K=1$	$K=5$
✗	✗	$70.68 \pm 0.61\%$	$83.36 \pm 0.45\%$
✗	✓	$70.93 \pm 0.66\%$	$83.76 \pm 0.43\%$
✓	✗	$71.18 \pm 0.63\%$	$83.70 \pm 0.45\%$
✓	✓	**72.35** $\pm 0.63\%$	**84.11** $\pm 0.43\%$

Table 2: The accuracy (%) of the classifiers with different training strategies of loss functions.

Regularization	$K=1$	$K=5$
Baseline	$64.02 \pm 0.70\%$	$82.58 \pm 0.45\%$
CutMix†	$64.83 \pm 0.72\%$	$80.89 \pm 0.51\%$
Mixup†	$64.93 \pm 0.69\%$	$81.55 \pm 0.47\%$
CutOut†	$64.85 \pm 0.68\%$	$81.53 \pm 0.48\%$
AFR(only $\boldsymbol{P}_S$)	$67.58 \pm 0.69\%$	$82.96 \pm 0.46\%$
AFR($\boldsymbol{f}_s + \bar{\boldsymbol{P}}_S$)	**72.35** $\pm 0.63\%$	**83.74** $\pm 0.43\%$

Table 3: The accuracy (%) of the classifiers trained with different regularization strategies. † is our implementation.

ness of the different components of attentive feature regularization (AFR). Meanwhile, we use the pre-trained word2vec (Li et al. 2019) to represent the labels with vectors. All experiments in the ablation study are conducted on 5-way-K-shot settings, where $K = 1$ or $K = 5$. We first evaluate the category selection and then introduce the experiments of different attention calculations and training strategies.

The influences of semantic selection (RQ1) The semantic selection is designed for feature regularization, thus we train the classifier Γ with only instance attention and $\mathcal{L}_{\mathrm{CE}}$ loss to validate its effects. In this ablation study, we conduct experiments with different β_s on $K = 1$ and $K = 5$, where β_s ranges from 1 to 64 (base categories of Mini-ImageNet). The results are shown in Figure 4. For comparison, we also plot the results without any operation at 0^{th} position ("Baseline"). First, both the introduced semantic selection and instance attention can improve the performance of the classifier. Moreover, the accuracy of using instance attention is better than that of utilizing the whole base categories (64^{th} position). Second, with the increase of category selection, the performances of the classifier increase first and then decrease. It's because introducing too many categories also bring more noise, which makes it hard to train the classifier. Therefore, we set $\beta_s = 3$ in our remaining experiments.

The effects of different attentions (RQ2) To evaluate the effectiveness of different attentions, we train four classifiers with or without attention operations, using only the $\mathcal{L}_{\mathrm{CE}}$ loss. The performance of each classifier was evaluated on $K = 1$ and $K = 5$, and the results are shown in Table 1. The results indicate that both instance and channel attention improve the classifier performance for the query samples. Compared to the classifier without employing any attention, the introduced instance attention and channel attention achieve nearly 6% accuracy improvements on the $K = 1$ experiment, respectively. More importantly, combining these attentions provides the best performance (the last row of Table 1), with over 7.5% improvement, which validates the effectiveness of our attention calculations.

The effectiveness of different losses (RQ3) In this ablation study, we train four classifiers with different loss functions, where the instance attention and the channel attention are applied in all cases. To balance the optimization process of these losses, we set $\mu_1 = 5$ and $\mu_2 = 20$ experientially and following (Li et al. 2022). The performances of four classifiers on $K = 1$ and $K = 5$ are shown in Table 2, which show that both $\mathcal{L}_{\mathrm{SC}}$ and $\mathcal{L}_{\mathrm{MSE}}$ contribute to the training procedure of the classifier. Moreover, combining these two losses further improves classification performance.

We also verify the effects of different regularizations in Table 3. The common regularizations, *i.e.* CutMix, Mixup, and CutOut, achieve slight improvement over the baseline in the 1-shot task but are harmful to accuracy in the 5-shot task. The classifier trained with the regularized features obtains over 3% improvements (**AFR**(only $\bar{\boldsymbol{P}}_S$)). Moreover, our AFR ($\boldsymbol{f}_s + \bar{\boldsymbol{P}}_S$) can further improve the performances.

Comparisons with Other Methods (RQ4)

We compare the performance of our method with the latest on the **Mini-ImageNet** and **Tiered-ImageNet** datasets. Table 4 shows the results which contain MatchingNets (Lee et al. 2019), ProtoNets (Snell, Swersky, and Zemel 2017), MixtFSL (Afrasiyabi, Lalonde, and Gagné 2021), RENet (Kang et al. 2021), DeepBDC (Xie et al. 2022), FeLMi (Roy et al. 2022), tSF (Lai et al. 2022), RankDNN (Guo et al. 2023), FRN (Wertheimer et al. 2021), BML (Zhou et al. 2021), FEAT (Ye et al. 2020), Label-Halluc (Jian and Torresani 2022), SEGA (Yang, Wang, and Chen 2022), IFSL (Yue et al. 2020), and LDRC (Yang, Liu, and Xu 2021). At the same time, we apply our approach to five recently proposed popular FSL methods, *i.e.*, Meta-Baseline, FRN, BML, FEAT, Label-Halluc, SEGA, and LRDC. We clearly observe that our approach consistently improves the classification performance in all settings, which is agnostic to the method, datasets, and pre-trained backbones. For different features extracted with various methods on Mini-ImageNet, we perform remarkable 6.61% accuracy improvements with the baseline ("Meta-Baseline + **AFR**") and obtain the best accuracy 74.57% with features from (Zhou et al. 2021) ("Label-Halluc + **AFR**") under $K = 1$. Generally, our AFR outperforms the compared methods by about 2% in accuracy for $K = 1$ and the improvements are generally greater in the 1-shot setting compared to the 5-shot setting. In the Tiered-ImageNet, we gain the 4.42% improvement ("BML + **AFR**") and achieve the best performance 89.59% ("LRDC + **AFR**") for $K = 1$ and $K = 5$, respectively.

To further demonstrate the effectiveness of our AFR, we conduct evaluations on the Meta-Dateset with $K = 1$ setting. The results are summarized in Table 5, including SimpleShot(Wang et al. 2019), ZN(Fei et al. 2021), and TCPR(Xu et al. 2022). We can see that our AFR exhibits strong adaptability to new data domains and achieves the best classification performance across several testing datasets. Notably, even compared to the transductive setting

Method	Backbone	Mini-ImageNet		Tiered-ImageNet	
		$K=1$	$K=5$	$K=1$	$K=5$
MatchingNets (NeurIPS16)	ResNet-12	$63.08 \pm 0.80\%$	$75.99 \pm 0.60\%$	$68.50 \pm 0.92\%$	$80.60 \pm 0.71\%$
ProtoNets (NeurIPS17)	ResNet-12	$60.37 \pm 0.83\%$	$78.02 \pm 0.57\%$	$65.65 \pm 0.92\%$	$83.40 \pm 0.65\%$
MixtFSL (ICCV21)	ResNet-12	$63.98 \pm 0.79\%$	$82.04 \pm 0.49\%$	$70.97 \pm 1.03\%$	$86.16 \pm 0.67\%$
RENet (ICCV21)	ResNet-12	$67.60 \pm 0.44\%$	$82.58 \pm 0.30\%$	$71.61 \pm 0.51\%$	$85.28 \pm 0.35\%$
DeepBDC (CVPR22)	ResNet-12	$67.34 \pm 0.43\%$	$84.46 \pm 0.28\%$	$72.34 \pm 0.49\%$	$87.31 \pm 0.32\%$
FeLMi (NeurIPS22)	ResNet-12	$67.47 \pm 0.78\%$	$86.08 \pm 0.44\%$	$71.63 \pm 0.89\%$	$87.01 \pm 0.55\%$
tSF(ECCV22)	ResNet-12	$69.74 \pm 0.47\%$	$83.91 \pm 0.30\%$	$71.89 \pm 0.50\%$	$85.49 \pm 0.35\%$
FEAT (CVPR20)	ResNet-12	$66.78 \pm 0.20\%$	$82.05 \pm 0.14\%$	$70.80 \pm 0.23\%$	$84.79 \pm 0.16\%$
FEAT + **AFR**	ResNet-12	**72.57** $\pm 0.62\%$	**85.06** $\pm 0.42\%$	**71.55** $\pm 0.74\%$	**87.64** $\pm 0.46\%$
Meta-Baseline (ICCV21)	ResNet-12	$63.17 \pm 0.23\%$	$79.26 \pm 0.17\%$	$68.62 \pm 0.27\%$	$83.74 \pm 0.18\%$
Meta-Baseline + **AFR**	ResNet-12	**69.78** $\pm 0.61\%$	**84.51** $\pm 0.41\%$	**69.66** $\pm 0.70\%$	**86.29** $\pm 0.48\%$
FRN (CVPR21)	ResNet-12	$66.45 \pm 0.19\%$	$82.83 \pm 0.13\%$	$71.16 \pm 0.22\%$	$86.01 \pm 0.15\%$
FRN + **AFR**	ResNet-12	**71.66** $\pm 0.56\%$	**84.75** $\pm 0.46\%$	**71.54** $\pm 0.71\%$	**87.35** $\pm 0.47\%$
BML (ICCV21)	ResNet-12	$67.04 \pm 0.63\%$	$83.63 \pm 0.29\%$	$68.99 \pm 0.50\%$	$85.49 \pm 0.34\%$
BML + **AFR**	ResNet-12	**73.84** $\pm 0.60\%$	**86.63** $\pm 0.41\%$	**73.41** $\pm 0.74\%$	**87.44** $\pm 0.48\%$
Label-Halluc (AAAI22)	ResNet-12	$68.28 \pm 0.77\%$	$86.54 \pm 0.46\%$	$73.34 \pm 1.25\%$	$87.68 \pm 0.83\%$
Label-Halluc + **AFR**	ResNet-12	**74.57** $\pm 0.58\%$	**87.30** $\pm 0.37\%$	**73.66** $\pm 0.66\%$	**89.15** $\pm 0.40\%$
SEGA (WACV22)	ResNet-12	$69.04 \pm 0.26\%$	$79.03 \pm 0.18\%$	$72.18 \pm 0.30\%$	$84.28 \pm 0.21\%$
SEGA + **AFR**	ResNet-12	**71.14** $\pm 0.60\%$	**84.26** $\pm 0.42\%$	**72.87** $\pm 0.45\%$	**85.26** $\pm 0.54\%$
IFSL (NeurIPS20)	WRN-28-10	$64.12 \pm 0.44\%$	$80.97 \pm 0.31\%$	$69.96 \pm 0.46\%$	$86.19 \pm 0.34\%$
tSF (ECCV22)	WRN-28-10	$70.23 \pm 0.46\%$	$84.55 \pm 0.29\%$	$74.87 \pm 0.49\%$	$88.05 \pm 0.32\%$
RankDNN (AAAI23)	WRN-28-10	$66.67 \pm 0.15\%$	$84.79 \pm 0.11\%$	$74.00 \pm 0.15\%$	$88.80 \pm 0.25\%$
FEAT (CVPR20)	WRN-28-10	$65.10 \pm 0.20\%$	$81.11 \pm 0.14\%$	$70.41 \pm 0.23\%$	$84.38 \pm 0.16\%$
FEAT + **AFR**	WRN-28-10	**71.76** $\pm 0.59\%$	**84.60** $\pm 0.42\%$	**71.74** $\pm 0.74\%$	**86.33** $\pm 0.53\%$
LRDC (ICLR21)	WRN-28-10	$68.57 \pm 0.55\%$	$82.88 \pm 0.42\%$	$74.38^{\dagger} \pm 0.93\%$	$88.12^{\dagger} \pm 0.59\%$
LRDC + **AFR**	WRN-28-10	**72.98** $\pm 0.62\%$	**86.91** $\pm 0.40\%$	**75.26** $\pm 0.67\%$	**89.59** $\pm 0.46\%$

Table 4: The accuracies (%) by different methods on the novel categories from Mini-ImageNet (Vinyals et al. 2016) and Tiered-ImageNet (Ren et al. 2018). † denotes our implementation.

Method	Testing Data Set							
	Mini-Test	CUB	Fungi	Omini	Sign	QDraw	Flower	DTD
SimpleShot (arXiv2019)	67.18%	49.68%	43.79%	78.19%	54.04%	54.50%	71.68%	51.19%
ZN (ICCV2021)	67.05%	48.15%	43.24%	78.80%	53.92%	52.86%	72.01%	52.20%
TCPR (NeurIPS 2022)	69.52%	53.83%	46.28%	80.88%	56.65%	57.31%	75.37%	54.38%
AFR	**72.98%**	**54.45%**	**47.93%**	**81.84%**	**60.12%**	**58.20%**	**76.11%**	**57.47%**

Table 5: The accuracies (%) by different methods on Meta-Dataset (Triantafillou et al. 2019) with $K = 1$. Sign and DTD denote Traffic Signs and Describable Textures dataset, respectively.

of TCPR, our approach gains more than 3% improvements on Traffic Signs and Describale Textures datasets.

Conclusion

In this paper, we have proposed attentive feature regularization named AFR to tackle the challenges in few-shot learning. Specifically, (1) The category selection based on semantic knowledge is employed to carefully constrain the features for regularization and helps avoid introducing unrelated noise into the training process. (2) Two attention calculations are designed to improve the complementarity of the features across the different categories and improve the channel discriminability of the regularized features. The extensive experiments have demonstrated the effectiveness of our proposed method, particularly in the 1-shot setting.

Note that the current usage of semantic relations is superficial. In our future work, we will focus on achieving a more robust feature regularization by incorporating additional techniques, such as GCN (Graph Convolutional Network) and GNN (Graph Neural Network), *et al.*, to further enhance the performance of the classifier.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants No. 62202439).

References

Afrasiyabi, A.; Lalonde, J.; and Gagné, C. 2021. Mixture-based Feature Space Learning for Few-shot Image Classification. In *ICCV*, 9021–9031.

Chen, Y.; Liu, Z.; Xu, H.; Darrell, T.; and Wang, X. 2021. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In *ICCV*, 9042–9051.

Chikontwe, P.; Kim, S.; and Park, S. H. 2022. CAD: Co-Adapting Discriminative Features for Improved Few-Shot Classification. In *CVPR*, 14534–14543.

Chou, H.; Chang, S.; Pan, J.; Wei, W.; and Juan, D. 2020. Remix: Rebalanced Mixup. In *ECCV Workshops (6)*, volume 12540, 95–110.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*, 3606–3613.

Deutsch, S.; Kolouri, S.; Kim, K.; Owechko, Y.; and Soatto, S. 2017. Zero Shot Learning via Multi-scale Manifold Regularization. In *CVPR*, 5292–5299.

Devries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *CoRR*, abs/1708.04552.

Fei, N.; Gao, Y.; Lu, Z.; and Xiang, T. 2021. Z-score normalization, hubness, and few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 142–151.

Fernandez-Fernandez, R.; Victores, J. G.; Estevez, D.; and Balaguer, C. 2019. Quick, stat!: A statistical analysis of the quick, draw! dataset. *arXiv preprint arXiv:1907.06417*.

Guo, H.; Mao, Y.; and Zhang, R. 2019. MixUp as Locally Linear Out-of-Manifold Regularization. In *AAAI*, 3714–3722.

Guo, Q.; Haotong, G.; Wei, X.; Fu, Y.; Yu, Y.; Zhang, W.; and Ge, W. 2023. RankDNN: Learning to Rank for Few-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 728–736.

Hariharan, B.; and Girshick, R. B. 2017. Low-Shot Visual Recognition by Shrinking and Hallucinating Features. In *ICCV*, 3037–3046.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cross Attention Network for Few-shot Classification. In *NeurIPS*, 4005–4016.

Hou, S.; Liu, X.; and Wang, Z. 2017. DualNet: Learn Complementary Features for Image Recognition. In *ICCV*, 502–510.

Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; and Igel, C. 2013. Detection of traffic signs in real-world images: The German traffic sign detection benchmark. In *IJCNN*, 1–8.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *CVPR*, 7132–7141.

Jian, Y.; and Torresani, L. 2022. Label hallucination for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7005–7014.

Kang, D.; Kwon, H.; Min, J.; and Cho, M. 2021. Relational Embedding for Few-Shot Classification. In *ICCV*, 8802–8813.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *NeurIPS*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Lai, J.; Yang, S.; Liu, W.; Zeng, Y.; Huang, Z.; Wu, W.; Liu, J.; Gao, B.; and Wang, C. 2022. tSF: Transformer-Based Semantic Filter for Few-Shot Learning. In *ECCV*.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 1332–1338.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-Learning With Differentiable Convex Optimization. In *CVPR*, 10657–10665.

Li, A.; Huang, W.; Lan, X.; Feng, J.; Li, Z.; and Wang, L. 2020. Boosting Few-Shot Learning With Adaptive Margin Loss. In *CVPR*, 12573–12581.

Li, A.; Luo, T.; Lu, Z.; Xiang, T.; and Wang, L. 2019. Large-Scale Few-Shot Learning: Knowledge Transfer With Class Hierarchy. In *CVPR*, 7212–7220.

Li, S.; Xia, X.; Ge, S.; and Liu, T. 2022. Selective-Supervised Contrastive Learning with Noisy Labels. In *CVPR*, 316–325.

Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; and Zhang, L. 2021. Learning a Few-shot Embedding Model with Contrastive Learning. In *AAAI*, 8635–8643.

Liu, N.; Zhao, Q.; Zhang, N.; Cheng, X.; and Zhu, J. 2019. Pose-Guided Complementary Features Learning for Amur Tiger Re-Identification. In *ICCV Workshops*, 286–293.

Lu, J.; Wang, S.; Zhang, X.; Hao, Y.; and He, X. 2023. Semantic-based Selection, Synthesis, and Supervision for Few-shot Learning. In *ACM Multimedia*, 3569–3578.

Luo, X.; Xu, J.; and Xu, Z. 2022. Channel Importance Matters in Few-Shot Image Classification. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 14542–14559. PMLR.

Nilsback, M.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*, 722–729.

Peng, Z.; Li, Z.; Zhang, J.; Li, Y.; Qi, G.; and Tang, J. 2019. Few-Shot Image Recognition With Knowledge Transfer. In *ICCV*, 441–449.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-Learning for Semi-Supervised Few-Shot Classification. In *ICLR*.

Rodríguez, P.; Laradji, I. H.; Drouin, A.; and Lacoste, A. 2020. Embedding Propagation: Smoother Manifold for Few-Shot Classification. In *ECCV*, 121–138.

Roy, A.; Shah, A.; Shah, K.; Dhar, P.; Cherian, A.; and Chellappa, R. 2022. FeLMi: Few shot Learning with hard Mixup. In *NeurIPS*.

Shi, C.; Wu, H.; and Wang, L. 2023. A Feature Complementary Attention Network Based on Adaptive Knowledge Filtering for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote. Sens.*, 61: 1–19.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In *NeurIPS*, 4077–4087.

Sulc, M.; Picek, L.; Matas, J.; Jeppesen, T. S.; and Heilmann-Clausen, J. 2020. Fungi Recognition: A Practical Use Case. In *WACV*, 2305–2313.

Triantafillou, E.; Zhu, T.; Dumoulin, V.; Lamblin, P.; Evci, U.; Xu, K.; Goroshin, R.; Gelada, C.; Swersky, K.; Manzagol, P.-A.; et al. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017a. Attention is All you Need. In *NeurIPS*, 5998–6008.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017b. Attention is All you Need. In *NeurIPS*, 5998–6008.

Velazquez, D.; Rodrıguez, P.; Gonfaus, J. M.; Roca, F. X.; and Gonzalez, J. 2022. A Closer Look at Embedding Propagation for Manifold Smoothing. *The Journal of Machine Learning Research*, 23: 1–27.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold Mixup: Better Representations by Interpolating Hidden States. In *ICML*, 6438–6447.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *NeurIPS*, 3630–3638.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, S.; Yue, J.; Liu, J.; Tian, Q.; and Wang, M. 2020. Large-Scale Few-Shot Learning via Multi-modal Knowledge Discovery. In *ECCV*, 718–734.

Wang, S.; Zhang, X.; Hao, Y.; Wang, C.; and He, X. 2022. Multi-directional Knowledge Transfer for Few-Shot Learning. In *ACM Multimedia*, 3993–4002.

Wang, Y.; Chao, W.-L.; Weinberger, K. Q.; and Van Der Maaten, L. 2019. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*.

Wang, Y.; Girshick, R. B.; Hebert, M.; and Hariharan, B. 2018. Low-Shot Learning From Imaginary Data. In *CVPR*, 7278–7286.

Wertheimer, D.; Tang, L.; Hariharan, B.; ; and and. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *CVPR*, 8012–8021.

Xie, J.; Long, F.; Lv, J.; Wang, Q.; and Li, P. 2022. Joint Distribution Matters: Deep Brownian Distance Covariance for Few-Shot Classification. In *CVPR*, 7962–7971.

Xu, J.; Luo, X.; Pan, X.; Li, Y.; Pei, W.; and Xu, Z. 2022. Alleviating the sample selection bias in few-shot learning by removing projection to the centroid. *Advances in Neural Information Processing Systems*, 35: 21073–21086.

Yang, F.; Wang, R.; and Chen, X. 2022. SEGA: Semantic guided attention on visual prototype for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1056–1066.

Yang, S.; Liu, L.; and Xu, M. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *ICLR*.

Ye, H.; Hu, H.; Zhan, D.; and Sha, F. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *CVPR*, 8805–8814.

Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X. 2020. Interventional Few-Shot Learning. In *NeurIPS*.

Yun, S.; Han, D.; Chun, S.; Oh, S. J.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *ICCV*, 6022–6031.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhou, Z.; Qiu, X.; Xie, J.; Wu, J.; and Zhang, C. 2021. Binocular Mutual Learning for Improving Few-shot Classification. In *ICCV*, 8382–8391.

Zhu, X.; Zhang, R.; He, B.; Zhou, A.; Wang, D.; Zhao, B.; and Gao, P. 2023. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. *arXiv preprint arXiv:2304.01195*.