Unleash the Power of Local Representations: Feature Calibration and Adaptive Metric for Few-Shot Learning

Shi Tang[1], **Chaoqun Chu**[1], **Guiming Luo**[1*], **Xinchen Ye**[2], **Zhiyi Xia**[1], **Haojie Li**[2]

[1]School of Software, Tsinghua University
[2]International School of Information Science&Engineering, Dalian University of Technology

Abstract

Recent metric-based few-shot learning (FSL) methods tend to adopt a set of local features instead of a global embedding to represent an instance, bridging base and novel class samples through potential common local features to improve novel-class generalization. However, due to **biased features** caused by treating local patches as base class samples during pretraining and a **non-adaptive metric** that cannot handle various local feature sets, existing methods are unable to take full advantage of local representations, leading to insufficient improvement in novel-class generalization. To address these issues, we investigate Feature Calibration (FC) and an Adaptive Metric (AM) to propose a novel method for FSL, namely FCAM. We treat local patches as "pseudo" novel class samples and generate soft labels capable of describing them to calibrate the biased features while fully exploiting their potential in improving novel-class generalization. Meanwhile, we employ the dual-Sinkhorn Divergence (dSD) with a designed Regulation Module (RM) to endow the metric with the flexibility to handle various local feature sets. Our method achieves new state-of-the-art on three popular benchmarks. Moreover, it exceeds state-of-the-art transductive and cross-modal methods in the fine-grained scenario.

Introduction

Few-shot learning aims to classify novel classes with limited examples based on a classifier constructed using abundant labeled instances from base classes. Among a series of proposed approaches (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Liu et al. 2022; Zhang et al. 2022; Finn, Abbeel, and Levine 2017; Schwartz et al. 2018; Li et al. 2019a), metric-based methods (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Liu et al. 2022; Zhang et al. 2022) are very elegant and promising. The main idea is to learn representations using deep networks and label the query sample by measuring its similarity to support samples.

Suffering from the low-data regimes and the inconsistency between training with base classes and testing with novel classes, FSL algorithms often struggle with poor novel-class generalization. Concretely, embeddings of congener samples are pushed far apart in the feature space (Zhang et al. 2020; Liu et al. 2022). Recent approaches (Zhang et al. 2020; Wertheimer, Tang, and Hariharan 2021; Liu et al. 2022; Zhang et al. 2022) try to solve it by using a set of local features to represent an instance instead of a global embedding, in the hope of providing transferrable information across categories through potential common local features between base and novel class samples. Generally, an instance is represented by a local feature set whose elements can be implemented as local feature vectors (Li et al. 2019b; Zhang et al. 2020; Liu et al. 2022; Zhang et al. 2022) or embeddings of patches cropped grid-like (Liu et al. 2022; Zhang et al. 2022) or randomly (Zhang et al. 2022). And then support-query pairs are measured by metrics capable of measuring two sets, e.g., accumulated cosine similarities between nearest neighbors (Li et al. 2019b), bidirectional random walk (Liu et al. 2022) or the Earth Mover's Distance (EMD) (Zhang et al. 2020, 2022).

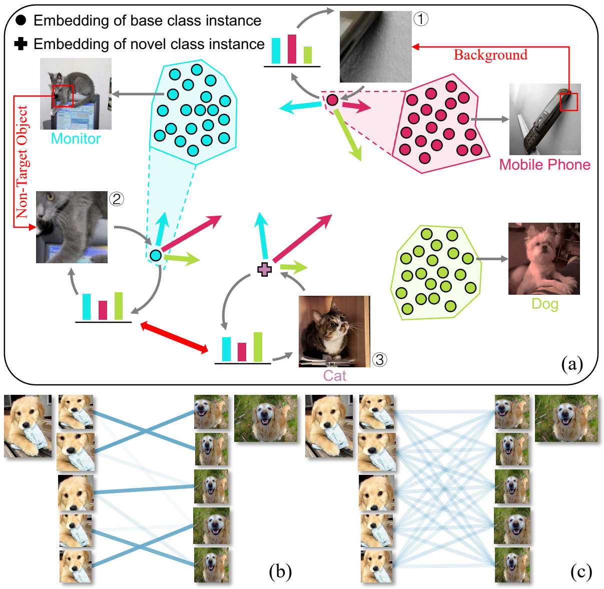
● Embedding of base class instance
✚ Embedding of novel class instance
①
Background
Monitor
Mobile Phone
Non-Target Object
②
Dog
③
Cat
(a)
(b)
(c)

Figure 1: (a) False supervision introduced by random cropping during pretraining leads to biased features, which can be calibrated by soft labels. And the matching flows between two sets of similar local patches using (b) EMD and (c) dSD.

[*]Corresponding author, gluo@tsinghua.edu.cn.
Preprint. Under review.

Despite the promising results, existing methods are unable to fully leverage local representations, limiting further improvement in novel-class generalization. The first reason is the biased features. Usually, a proxy task of classifying all the base classes is adopted to pre-train the encoder before episodic training (Chen et al. 2021b; Liu et al. 2022; Zhang et al. 2022; Hu et al. 2022), during which random cropping is often used for augmentation following the conventional paradigm. However, the main content of a patch may be the background (Figure 1 (a) ①) or non-target objects (Figure 1 (a) ②) whose semantic differs from the uncropped raw image. The ground-truth label can thus be false supervision, leading to biased features incapable of describing input images accurately. Although possible false supervision can be avoided by removing random cropping, it will reduce the diversity of training data and cause performance degradation instead. Besides, due to the semantic difference, the cropped patches can be thought of as "pseudo" novel class samples providing class-level diversity, which can be used to prevent the network from overfitting to base classes. Moreover, non-target objects (Figure 1 (a) ②) may be related to novel classes (Figure 1 (a) ③), which can be utilized to warm up the encoder, narrowing the gap between training and testing.

The second reason is the non-adaptive metric. With the form of the well-studied optimal transport (OT) problem, EMD exhibits great superiority in measuring two local feature sets (Zhang et al. 2022). However, it lacks the ability to handle sets consisting of similar local features, making it not adaptive enough to measure various local feature sets, especially when the local features are embeddings of randomly cropped patches. Specifically, the optimum transport matrix is usually solved on a vertex of the transport polytope (Cuturi 2013), resulting in a sparse transport matrix. As a consequence, EMD tries to match a patch with a few "most" similar opposite patches even when the opposite patches are homogeneous, as shown in Figure 1 (b). For a more accurate metric, a smoother transport matrix that allows "one-to-many" matching is desired under such circumstances to utilize the opposite patches comprehensively and reduce the dependency on a few opposite patches.

In this paper, we investigate Feature Calibration and an Adaptive Metric to unleash the power of local representations for few-shot image classification. For the biased features, we generate soft labels to supervise the learning of local patches during pretraining in the form of self-distillation, which not only corrects false supervision but also regularize and adapt the encoder to the test scenario in advance. In addition, by decomposing the classical KL-Divergence commonly used in self-distillation, we find its inherent weighting scheme unsuitable for distilling FSL networks. Therefore, we propose Smoothed KL-Divergence (SKD) with a smoother weighting scheme more suitable for the task of FSL. For the non-adaptive metric, we employ the dual-Sinkhorn Divergence (Cuturi 2013) to endow the algorithm with the ability to handle sets consisting of similar local features (Figure 1 (c)). With a designed Regulation Module, we implement an Adaptive Metric by self-adaptively controlling the transport matrix's smoothness. Our method achieves new state-of-the-art on three popular benchmarks. Moreover, it exceeds state-of-the-art transductive and cross-modal methods in the fine-grained scenario. In summary, our main contributions are as follows:

- We investigate Feature Calibration and an Adaptive Metric to unleash the power of local representations in improving novel-class generalization for FSL.
- We propose a novel pretraining paradigm for FSL, along with a designed Smoothed KL-Divergence more suitable for distilling FSL networks, to calibrate the biased features towards the test scenario.
- We introduce an Adaptive Metric capable of handling various sets adaptively with the dual-Sinkhorn Divergence and a constructed Regulation Module.

Related Work

Metric-based few-shot learning. The literature exhibits significant diversity in the area of FSL. Under the meta-learning framework, metric-based methods advocate to meta-learn a representation expected to be generalizable across categories with a predefined (Snell, Swersky, and Zemel 2017; Liu et al. 2022; Zhang et al. 2020, 2022) or also meta-learned (Sung et al. 2018) metric as the classifier. For example, Snell *et al.* (Snell, Swersky, and Zemel 2017) average the embeddings of congener support samples as the class prototype and leverages the Euclidean distance for classification. Sung *et al.* (Sung et al. 2018) replace the metric with a learnable module to introduce nonlinearity. To avoid congener image-level embeddings from being pushed far apart by the significant intra-class variations, recent approaches tend to employ a set of local features to represent an instance instead of a global embedding. Accordingly, the metric between global features becomes a metric between local feature sets. Zhang *et al.* (Zhang et al. 2022) propose three methods for constructing local feature sets and employs EMD for measuring two sets. Liu *et al.* (Liu et al. 2022) adopt bidirectionally random walk for measurement to affiliate two feature sets in a bidirectional paradigm. Our method is also based on local features; the main differences lie in the encoder and the metric. We calibrate the encoder towards the test scenario for higher quality local features and adopt dSD with RM to overcome the shortcomings of EMD.

Self-distillation. First proposed for model compression (Buciluundefined, Caruana, and Niculescu-Mizil 2006; Ba and Caruana 2014; Hinton, Vinyals, and Dean 2015), knowledge distillation aims at transferring "knowledge", such as logits (Hinton, Vinyals, and Dean 2015) or intermediate features (Yim et al. 2017; Kim, Park, and Kwak 2018; Heo et al. 2019), from a high-capability teacher model to a lightweight student network. As a special case when the teacher and student architectures are identical, self-distillation has been consistently observed to achieve higher accuracy (Furlanello et al. 2018). Zhang *et al.* (Zhang and Sabuncu 2020) relate self-distillation with label smoothing, a commonly-used regularization technique to prevent models from being overconfident. Inspired by the fact that the smoothed soft labels can be used to describe "pseudo" novel class samples, we

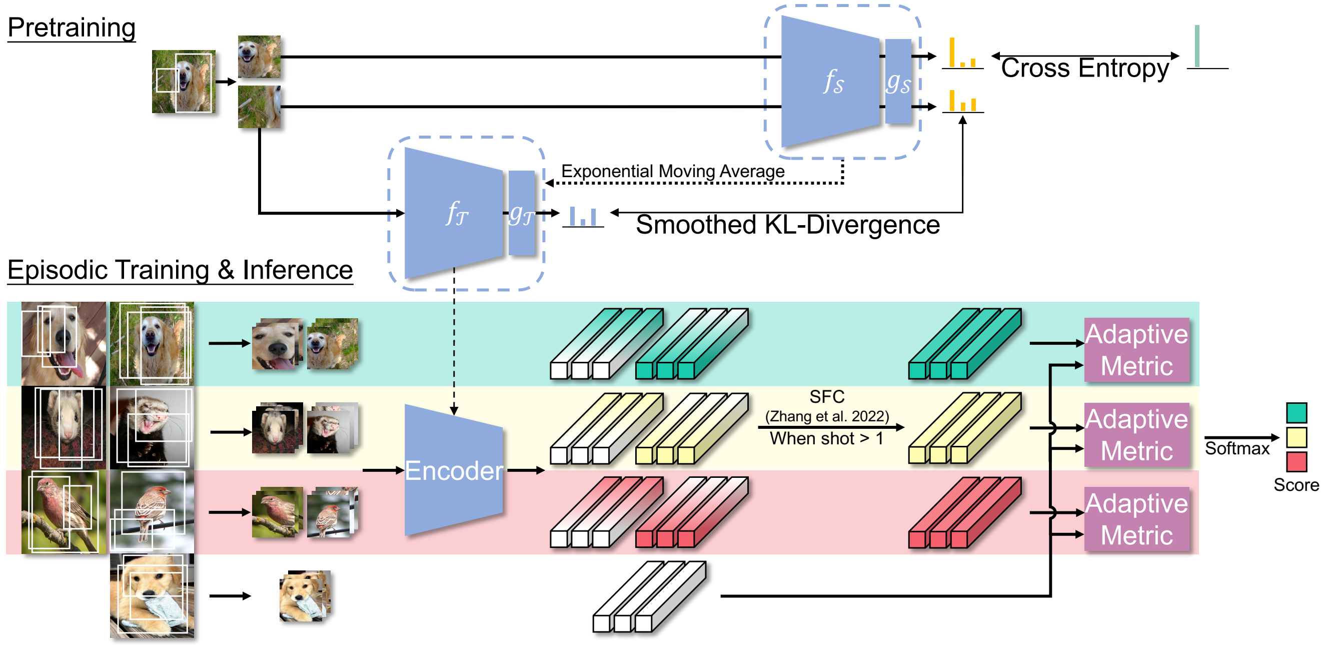

Figure 2: Overview of our framework (3-way 2-shot as an example).

find that the regularization effect of self-distillation is also beneficial in improving novel-class generalization.

Optimal transport distances. The distances based on the well-studied OT problem are very powerful for probability measures. EMD was first proposed for image retrieval (Rubner, Guibas, and Tomasi 1997) and exhibited excellent performance. Cuturi (Cuturi 2013) regularizes the OT problem with an entropic term, which greatly improves the computing efficiency and defines a distance with a natural prior on the transport matrix: everything should be homogeneous in the absence of a cost. As an essential prior that EMD lacks, it is crucial in dealing with various local feature sets.

Methods

As illustrated in Figure 2, we integrate our method into the "pretraining + episodic training" paradigm. In the pretraining stage, the biased features are calibrated through self-distillation with SKD; and in the episodic training stage, the classification is based on dSD with RM. In the following, we first briefly introduce some preliminary concepts and then present our method in detail.

Preliminary

Given a labeled dataset $D_{base} = \{(x_i^b, y_i^b)\}_{i=1}^{N_{base}}$ composed of n_c base classes, the goal of FSL is to handle tasks consisting of novel classes. Generally, an N-way K-shot Q-query task is described by a task-specific pair of datasets $(D_{support}, D_{query})$. Containing N classes with K samples per class, $D_{support} = \{(x_i^s, y_i^s)\}_{i=1}^{NK}$ provides examples for reference, according to which we need to assign labels for $D_{query} = \{x_i^q\}_{i=1}^{NQ}$ that contains samples from the same N classes with Q samples per class.

Feature Calibration

Feature Calibration with self-distillation Different from existing methods that supervise the learning of cropped local patches with ground-truth hard labels during pretraining, we advocate taking soft labels into account as well. Indicating the probability of a patch belonging to each base class, they implicitly take base class prototypes as the manifold base to represent patches, which can properly describe the background or non-target objects[1], making it possible to use these "pseudo" novel class samples for regularization while correcting false supervision. Moreover, soft labels connect potential novel classes with related non-target objects through similar distributions (Figure 1 (a) ②③), making the learning of these patches a pre-search for areas suitable for potential novel classes in the feature space, which adapts the encoder to potential test scenario in advance.

Therefore, based on the proxy task of standard classification on D_{base}, we propose a novel paradigm for pretraining the encoder, calibrating the biased features while fully exploiting the potential of local patches in improving novel-class generalization in the form of self-distillation. As shown in Figure 2, the pretraining stage involves two structurally identical networks, i.e., a student network $\phi_{\mathcal{S}} = f_{\mathcal{S}} \circ g_{\mathcal{S}}$ and a teacher network $\phi_{\mathcal{T}} = f_{\mathcal{T}} \circ g_{\mathcal{T}}$, with $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ being their respective encoders, and $g_{\mathcal{S}}$ and $g_{\mathcal{T}}$ being their respective last linear layers. Given a sample x in D_{base}, a set of patches $\{\hat{x}_i | i = 1, 2, ..., n_p\}$ can be obtained by random cropping (with resize and flip). The first element $\hat{x}_1$ is reserved for

[1] Embeddings in the manifold space can be represented linearly or nonlinearly by the manifold base (Candès et al. 2011; Yue et al. 2020).

learning standard classification using the cross-entropy loss:

$$\mathcal{L}_{CE} = -\mathbf{y}^{\top} \log \left(\sigma(\phi_{\mathcal{S}}(\hat{x}_1))\right), \tag{1}$$

where σ denotes the softmax function and $\mathbf{y}$ is the label of x which is a one-hot vector. And the remaining $n_p - 1$ patches are used for distillation where the teacher for generating soft labels is momentum updated. In detail, denote the parameters of $\phi_{\mathcal{T}}$ as $\theta_{\mathcal{T}}$ and those of $\phi_{\mathcal{S}}$ as $\theta_{\mathcal{S}}$, for the i-th iteration, $\theta_{\mathcal{T}}$ is updated by (Tarvainen and Valpola 2017):

$$\theta_{\mathcal{T}}^{i} \leftarrow m\theta_{\mathcal{T}}^{i-1} + (1-m)\theta_{\mathcal{S}}^{i}, \tag{2}$$

where $m \in [0, 1)$ is a momentum coefficient. As an exponential moving average of the student, the teacher evolves more smoothly, which ensures the stability of the generated soft labels (He et al. 2020; Grill et al. 2020).

Smoothed KL-Divergence Usually, KL-Divergence is a common choice of the loss function in logit-level knowledge distillation. However, we find it not suitable for distilling networks for FSL, which we elaborate on by reformulating it. Let $\mathbf{z} = [z_1, z_2, ..., z_{n_c}] \in \mathbb{R}^{1\times n_c}$ denote the network output for a local patch where z_i represents the logit of the i-th base class. Considering a process of continuous binary classification where each time we focus on the distinction between the sample belonging to one class and belonging to the remaining classes as illustrated in Figure 3, the probabilities of the i-th binary classification $\mathbf{b}_i = [p_i, p_{\neg i}]$ can be obtained by:

$$p_i = \frac{\exp(z_i)}{\sum_{j=i}^{n_c} \exp(z_j)}, \quad p_{\neg i} = \frac{\sum_{k=i+1}^{n_c} \exp(z_k)}{\sum_{j=i}^{n_c} \exp(z_j)}. \tag{3}$$

Note that this is a process without replacement, i.e., the computation of $\mathbf{b}_i$ only involves the logits of class i-n_c. With the above notations, we can define the i-th binary classification probabilities of the teacher $\mathbf{b}_i^{\mathcal{T}}$ and the student $\mathbf{b}_i^{\mathcal{S}}$ using their respective outputs $\mathbf{z}^{\mathcal{T}}$ and $\mathbf{z}^{\mathcal{S}}$. And the classical KL-Divergence can be reformulated as (proof of Eq. (4) is presented in the supplementary material):

$$KL(\sigma(\mathbf{z}^{\mathcal{T}})||\sigma(\mathbf{z}^{\mathcal{S}})) = \sum_{i=1}^{n_c-1} \mathbf{w}_i \cdot KL(\mathbf{b}_i^{\mathcal{T}}||\mathbf{b}_i^{\mathcal{S}}), \quad \text{where } \mathbf{w}_i = \frac{\sum_{k=i}^{n_c} \exp(z_k^{\mathcal{T}})}{\sum_{j=1}^{n_c} \exp(z_j^{\mathcal{T}})}, \tag{4}$$

which intuitively demonstrates how KL-Divergence decomposes the problem of measuring two probability distributions of classification, i.e., by constantly measuring $\mathbf{b}_i$. Since continuous binary classification is a process without replacement, the number of remaining classes to be considered (class cardinality) differs for different i. Therefore, we consider a comparable form which normalizes $\mathbf{w}_i$ with the class cardinality $n_c - i + 1$:

$$\tilde{\mathbf{w}}_i = \frac{\sum_{k=i}^{n_c} \exp(z_k^{\mathcal{T}})}{(n_c - i + 1)\sum_{j=1}^{n_c} \exp(z_j^{\mathcal{T}})}. \tag{5}$$

With $\tilde{\mathbf{w}}_i$ coupled with $\mathbf{z}$, the less similar the teacher thinks the sample is to class i-n_c, the less important the alignment of $\mathbf{b}_i$. In the context of FSL, a sample does not necessarily belong to any base class. As a result, this weighting scheme introduces a false prior which emphasizes the penalties related to base classes similar to the sample more. Ideally, the probabilities of each binary classification should be valued equally because each of them provides vital information about a specific base class prototype used to describe local patches. Noticing that $\tilde{\mathbf{w}}_i = \sum_{k=i}^{n_c} \sigma_k(\mathbf{z}^{\mathcal{T}})/(n_c - i + 1)$, the weights can be smoothed by adjusting the distribution of $\sigma(\mathbf{z}^{\mathcal{T}})$. Following this trail, we introduce a temperature coefficient T to alter the inherent weighting scheme of KL-Divergence, i.e., $\tilde{\mathbf{w}}_i(T) = \sum_{k=i}^{n_c} \sigma_k(\mathbf{z}^{\mathcal{T}}/T)/(n_c - i + 1)$. Furthermore, we discover that the difference between the weights of two different binary classification processes $\tilde{\mathbf{w}}_\alpha(T)$ and $\tilde{\mathbf{w}}_\beta(T)$ ($\alpha \neq \beta$) vanishes with an extremely high temperature:

$$\begin{aligned} &\lim_{T\to\infty} |\tilde{\mathbf{w}}_\alpha(T) - \tilde{\mathbf{w}}_\beta(T)| \\ &= \lim_{T\to\infty} \left| \frac{\sum_{k_1=\alpha}^{n_c} \sigma_{k_1}(\mathbf{z}^{\mathcal{T}}/T)}{n_c - \alpha + 1} - \frac{\sum_{k_2=\beta}^{n_c} \sigma_{k_2}(\mathbf{z}^{\mathcal{T}}/T)}{n_c - \beta + 1} \right| = 0, \end{aligned} \tag{6}$$

according to which we introduce a temperature $T \to \infty$ into $\mathbf{w}_i$ to derive a smoothed weighting scheme suitable for FSL:

$$\mathbf{w}_i' = \lim_{T\to\infty} \frac{\sum_{k=i}^{n_c} \exp(z_k^{\mathcal{T}}/T)}{\sum_{j=1}^{n_c} \exp(z_j^{\mathcal{T}}/T)} = \frac{n_c - i + 1}{n_c}. \tag{7}$$

With a more rational weighting scheme, we define Smoothed KL-Divergence that is used to compute the distillation loss $\mathcal{L}_{SKD}$:

$$SKD(\sigma(\mathbf{z}^{\mathcal{T}})||\sigma(\mathbf{z}^{\mathcal{S}})) \coloneqq \sum_{i=1}^{n_c-1} \mathbf{w}_i' \cdot KL(\mathbf{b}_i^{\mathcal{T}}||\mathbf{b}_i^{\mathcal{S}}), \tag{8}$$

$$\mathcal{L}_{SKD} = \frac{1}{n_p - 1} \sum_{i=2}^{n_p} SKD(\sigma(\phi_T(\hat{x}_i))||\sigma(\phi_S(\hat{x}_i))). \tag{9}$$

And with a weight λ, $\mathcal{L}_{SKD}$ is combined with $\mathcal{L}_{CE}$ to form the total loss for pretraining:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{SKD}. \tag{10}$$

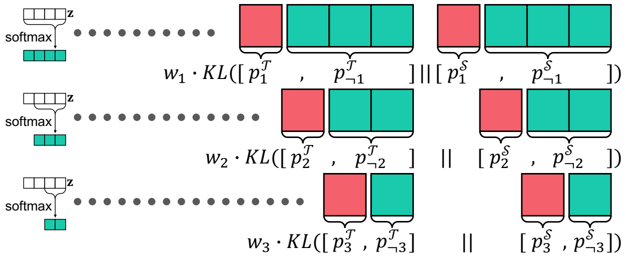
z
softmax
$w_1 \cdot KL([p_1^{\mathcal{T}}, p_{\neg 1}^{\mathcal{T}}]||[p_1^{\mathcal{S}}, p_{\neg 1}^{\mathcal{S}}])$
z
softmax
$w_2 \cdot KL([p_2^{\mathcal{T}}, p_{\neg 2}^{\mathcal{T}}] || [p_2^{\mathcal{S}}, p_{\neg 2}^{\mathcal{S}}])$
z
softmax
$w_3 \cdot KL([p_3^{\mathcal{T}}, p_{\neg 3}^{\mathcal{T}}] || [p_3^{\mathcal{S}}, p_{\neg 3}^{\mathcal{S}}])$

Figure 3: Illustration of the continuous binary classification process corresponding to the reformulation of KL-Divergence.

Adaptive Metric

After pretraining, $f_{\mathcal{T}}$ will be used as the feature extractor for further episodic training where we propose to employ dSD

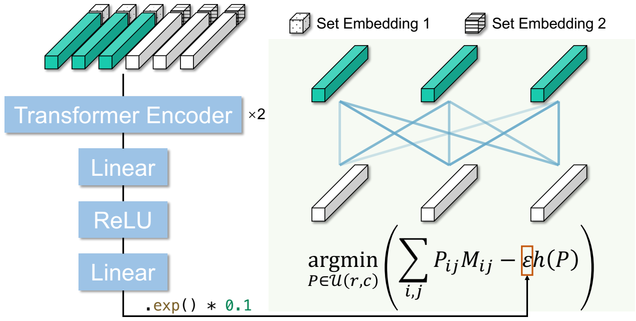

Figure 4: The proposed Adaptive Metric based on the dual-Sinkhorn Divergence and the Regulation Module.

with RM as shown in Figure 4 to produce a similarity score for each support-query pair.[2]

The dual-Sinkhorn Divergence for few-shot classification Following (Zhang et al. 2022), we generate local representations by random cropping (with resize and flip) and formulate the problem of measuring support-query pairs as an OT problem. Specifically, given a support-query pair, two sets of local patches $U = \{u_i | i = 1, 2, ..., n\}, V = \{v_j | j = 1, 2, ..., n\}$ can be generated. The OT problem considers a hypothetical process of transporting goods from suppliers U to demanders V. Corresponding to the importance of each patch in a set, the total supply (demand) units of the i-th supplier $\mathbf{r}_i$ (demander $\mathbf{c}_i$) can be obtained by the cross-reference mechanism (Zhang et al. 2022) followed by normalization to make both sides have the same total units for matching:

$$\hat{\mathbf{r}}_i = \max\Big\{\frac{1}{n}\sum_{j=1}^{n} f_{\mathcal{T}}(u_i)^\top f_{\mathcal{T}}(v_j), 0\Big\}, \quad \mathbf{r}_i = \frac{n\hat{\mathbf{r}}_i}{\sum_{j=1}^{n}\hat{\mathbf{r}}_j}, \quad (11)$$

$$\hat{\mathbf{c}}_i = \max\Big\{\frac{1}{n}\sum_{j=1}^{n} f_{\mathcal{T}}(v_i)^\top f_{\mathcal{T}}(u_j), 0\Big\}, \quad \mathbf{c}_i = \frac{n\hat{\mathbf{c}}_i}{\sum_{j=1}^{n}\hat{\mathbf{c}}_j}. \quad (12)$$

In addition, a cost matrix M whose elements denote the cost to transport a unit from node u_i to v_j is defined as:

$$M_{ij} = 1 - \frac{f_{\mathcal{T}}(u_i)^\top f_{\mathcal{T}}(v_j)}{\|f_{\mathcal{T}}(u_i)\|\|f_{\mathcal{T}}(v_j)\|}. \quad (13)$$

Based on the above notations, the goal of the OT problem is to find a transportation plan with the lowest total cost from a set of valid plans $\mathcal{U}(\mathbf{r}, \mathbf{c}) = \{P \in \mathbb{R}_{+}^{n\times n} | P\mathbf{1}_n = \mathbf{r}, P^\top \mathbf{1}_n = \mathbf{c}\}$. Hence, EMD can be obtained by solving $\arg\min_{P\in\mathcal{U}(\mathbf{r},\mathbf{c})} \sum_{i,j} P_{ij}M_{ij}$ (Rubner, Guibas, and Tomasi 1997). Different from EMD, the dual-Sinkhorn Divergence (Cuturi 2013) encourages a smoother transport matrix by introducing an entropic regularization:

$$P^{\varepsilon} = \mathop{\arg\min}_{P\in\mathcal{U}(\mathbf{r},\mathbf{c})}\Big(\sum_{i,j} P_{ij}M_{ij} - \varepsilon h(P)\Big), \quad (14)$$

where $h(P) = -\sum_{i,j} P_{ij}\log P_{ij}$ is the information entropy of P and $\varepsilon \in (0, \infty)$ serves as an adjustment coefficient. $h(P)$ reflects the smoothness of P, the higher the entropy, the smoother the matrix. By introducing the entropic regularization term into the optimization objective, dSD not only endows our method with the ability to handle sets consisting of similar local patches but also makes it possible to solve the transport problem faster as it becomes a strictly convex problem (Cuturi 2013) that can be solved with the Sinkhorn-Knopp algorithm (Sinkhorn and Knopp 1967) efficiently. With the solved P^{ε}, the similarity of a support-query pair can be obtained and be used to compute the classification score:

$$s(U, V) = \sum_{i=1}^{n}\sum_{j=1}^{n}(1 - M_{ij})P_{ij}^{\varepsilon}. \quad (15)$$

Regulation Module According to Eq. (14), ε can be used to control the smoothness of $P\varepsilon$. By making ε higher, P^{ε} will be smoother, and as ε goes to zero, it will be sparser, with the solution close to EMD. Therefore, based on the idea of making ε a learnable parameter, we design an RM to control the smoothness of the transport matrix adaptively according to the characteristics of the local feature sets. Intuitively, the smoothness of the transport matrix should be conditioned on the relationship of the local features (similar features come with similar local patches where a smooth transport matrix is expected). Therefore, we take the embedded local features as input and construct a predictor based on the Transformer encoder (Vaswani et al. 2017), considering that its inductive bias suits the task of modeling the relationship between local features very well. As shown in Figure 4, the input embeddings are constructed by concatenating the local feature with a 16 dimensional learnable set embedding indicating which set the local patch is from, i.e., support or query. Followed by an exponential function, the output serves as a scaling factor to adjust ε based on the default value of 0.1.

Experiments

Datasets. The experiments are conducted on three popular benchmarks: (1) ***mini*ImageNet** (Vinyals et al. 2016) is a subset of ImageNet (Russakovsky et al. 2015) that contains 100 classes with 600 images per class. The 100 classes are divided into 64/16/20 for train/val/test respectively; (2) ***tiered*ImageNet** (Ren et al. 2018) is also a subset of ImageNet (Russakovsky et al. 2015) that includes 608 classes from 34 super-classes. The super-classes are split into 20/6/8 for train/val/test respectively; (3) **CUB-200-2011** (Wah et al. 2011) contains 200 bird categories with 11,788 images, which represents a fine-grained scenario. Following the splits in (Ye et al. 2020), the 200 classes are divided into 100/50/50 for train/val/test respectively.

Backbone. For the backbone, we employ *ResNet12* as many previous works. With the dimension of the embedded features and the set embeddings being 640 and 16, respectively, we set $d_{model} = 656$, $d_{feedforward} = 1280$ and $n_{head} = 16$ for the 2-layer Transformer encoder in our RM.

[2] For cases where shot $K > 1$, all the local features of the support set are used to learn a prototype feature set with the structured FC layer (Zhang et al. 2022) and the latter process is the same as the 1-shot case.

Training details. In the pretraining stage, we set $n_p = 4$ and $m = 0.999$. $\mathcal{L}_{SKD}$ will not be used during early epochs to ensure the teacher has well-converged before being used to generate soft labels. In the episodic training stage, each epoch involves 50 iterations with a batch size of 4. We set $n = 25$, and the patches are resized to 84×84 before being embedded. RM is first pre-trained for 100 epochs with the encoder's parameters fixed, in which the learning rate starts from $1e$-3 and decays by 0.1 at epoch 60 and 90. Then, all the parameters will be optimized jointly for another 100 epochs. More detailed training settings are described in the supplementary material.

Comparison with State-of-the-art Methods

For general few-shot classification, we compare our method with the state-of-the-art methods in Table 1. Our method outperforms the state-of-the-art methods on all the settings and even achieves higher performance than methods with bigger backbones, achieving new state-of-the-art. For fine-grained few-shot classification, we compare our method with the state-of-the-art methods in Table 2. Benefit from higher quality local features, the discriminative regions can be depicted more accurately, resulting in significant improvement against other methods, i.e., $\mathbf{4.80\%}$ and $\mathbf{3.03\%}$ for 1-shot and 5-shot respectively against previous state-of-the-art method (Liu et al. 2022). In particular, our method even outperforms state-of-the-art transductive (Chen et al. 2021a, 2022) and cross-modal (Huang et al. 2021; Chen et al. 2022) methods, shedding some light on how much the poor local representations can degrade the performance in the fine-grained scenario.

Ablation Study

To begin with, a coarse-scale ablation is presented in Table 3. The baseline follows the traditional pretraining paradigm that uses only $\mathcal{L}_{CE}$ for supervision and employs EMD as the metric. With both FC and AM outperforming the baseline and achieving optimal results when used together, their respective effectiveness can be validated. Furthermore, we conduct a more detailed analysis below.

Feature Calibration improves novel-class generalization To demonstrate that Feature Calibration improves novel-class generalization, we visualize the 1-shot test accuracy change during self-distillation in Figure 5. We first pre-train the network to its highest validation accuracy with only $\mathcal{L}_{CE}$ to ensure the quality of the teacher and exclude the influence of hard label supervision on accuracy improvement during self-distillation. We observe a continuous improvement in test accuracy during distillation. In the case of our method ($T \to \infty$), the 1-shot accuracy is boosted from 70.20% to 77.41%, demonstrating the effectiveness of Feature Calibration in improving novel-class generalization and suggesting how severe the power of local representations is limited. In addition, the feature distributions visualized in Figure 6 also illustrate that Feature Calibration results in better clusters for novel classes.

SKD is more suitable for Feature Calibration We compare different temperature settings in Figure 5. It can be seen that the temperature, i.e., the weighting scheme, affects the process of Feature Calibration. A general trend that better test accuracy comes with higher temperature can be observed, and the setting corresponding to our SKD, i.e., $T \to \infty$, constantly outperforms other settings. Furthermore, SKD yields better final performance than the classical KL-Divergence (CKD) as shown in Table 4. Both the above experiments demonstrate the importance of a smoother weighting scheme in Feature Calibration.

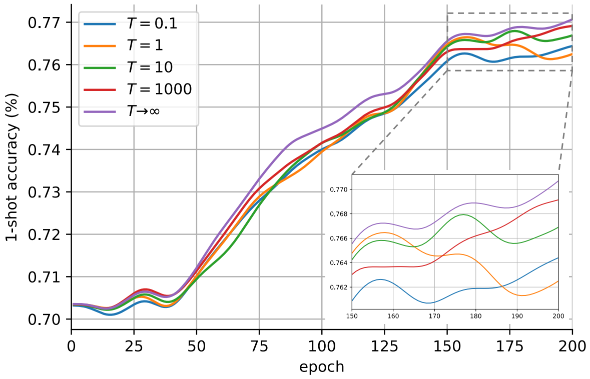

Figure 5: Gaussian smoothed 1-shot test accuracy curves on CUB-200-2011 during self-distillation, with different temperatures to adjust the weighting scheme of the classical KL-Divergence. The results of the same 1000 tasks are averaged for each data point.

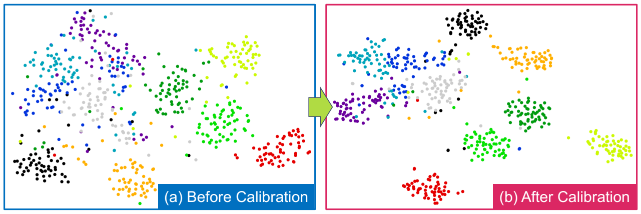

Figure 6: The t-sne visualization (van der Maaten and Hinton 2008) of novel class samples embedded by encoders trained (a) without and (b) with Feature Calibration.

The dual Sinkhorn-Divergence handles sets consisting of similar local patches For sets consisting of similar local patches, the transport matrix solved by EMD (Figure 7 (a)) is very sparse, which tries to match a patch with few "most" similar opposite patches. In contrast, dSD (Figure 7 (b)) generates a smoother transport matrix, which enables a comprehensive utilization of opposite patches and reduces the dependency on a few opposite patches by allowing "one-to-many" matching.

RM enables Adaptive Metric For sets consisting of similar local patches, RM produces a relatively larger ε, resulting in a smoother transport matrix (Figure 7 (b)). While for sets consisting of dissimilar local patches, a relatively smaller ε is predicted, making the transport matrix moderately sparse (Figure 7 (c)). Quantitative results of whether using RM to

Method	Backbone	*mini*ImageNet		*tiered*ImageNet	
		1-shot	5-shot	1-shot	5-shot
MatchNet† (Vinyals et al. 2016)	*ResNet12*	63.08 ± 0.80	75.99 ± 0.60	68.50 ± 0.92	80.60 ± 0.71
ProtoNet† (Snell, Swersky, and Zemel 2017)	*ResNet12*	60.37 ± 0.83	78.02 ± 0.57	65.65 ± 0.92	83.40 ± 0.65
TADAM (Oreshkin, Rodríguez López, and Lacoste 2018)	*ResNet12*	58.50 ± 0.30	76.70 ± 0.30	-	-
FEAT (Ye et al. 2020)	*ResNet12*	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
DeepEMD (Zhang et al. 2020)	*ResNet12*	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
Meta-Baseline (Chen et al. 2021b)	*ResNet12*	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.74 ± 0.18
FRN (Wertheimer, Tang, and Hariharan 2021)	*ResNet12*	66.45 ± 0.19	82.83 ± 0.13	72.06 ± 0.22	86.89 ± 0.14
PAL (Ma et al. 2021)	*ResNet12*	$\underline{69.37 \pm 0.64}$	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47
MCL (Liu et al. 2022)	*ResNet12*	69.31 ± 0.20	$\underline{85.11 \pm 0.20}$	73.62 ± 0.20	86.29 ± 0.20
DeepEMD v2 (Zhang et al. 2022)	*ResNet12*	68.77 ± 0.29	84.13 ± 0.53	$\underline{74.29 \pm 0.32}$	$\underline{87.08 \pm 0.60}$
Centroid Alignment‡ (Afrasiyabi, Lalonde, and Gagn'e 2020)	*WRN-28-10*	65.92 ± 0.60	82.85 ± 0.55	74.40 ± 0.68	86.61 ± 0.59
Oblique Manifold‡ (Qi et al. 2021)	*ResNet18*	63.98 ± 0.29	82.47 ± 0.44	70.50 ± 0.31	86.71 ± 0.49
FewTURE‡ (Hiller et al. 2022)	*ViT-Small*	68.02 ± 0.88	84.51 ± 0.53	72.96 ± 0.92	86.43 ± 0.67
FCAM (ours)	*ResNet12*	$\mathbf{70.20 \pm 0.28}$	$\mathbf{85.61 \pm 0.52}$	$\mathbf{74.90 \pm 0.32}$	$\mathbf{88.04 \pm 0.58}$

† results are reported in (Zhang et al. 2022). ‡ methods with bigger backbones. The second best results are underlined.

Table 1: Comparison to the state-of-the-art methods on *mini*ImageNet and *tiered*ImageNet, ordered chronologically. Average 5-way 1-shot and 5-way 5-shot accuracy (%) with 95% confidence intervals.

Method	Backbone	CUB-200-2011	
		1-shot	5-shot
MatchNet† (Vinyals et al. 2016)	*ResNet12*	71.87 ± 0.85	85.08 ± 0.57
ProtoNet† (Snell, Swersky, and Zemel 2017)	*ResNet12*	66.09 ± 0.92	82.50 ± 0.58
DeepEMD (Zhang et al. 2020)	*ResNet12*	75.65 ± 0.83	88.69 ± 0.50
FRN♯ (Wertheimer, Tang, and Hariharan 2021)	*ResNet12*	78.86 ± 0.28	$\underline{90.48 \pm 0.16}$
MCL♯ (Liu et al. 2022)	*ResNet12*	$\underline{79.39 \pm 0.29}$	$\underline{90.48 \pm 0.49}$
DeepEMD v2 (Zhang et al. 2022)	*ResNet12*	79.27 ± 0.29	89.80 ± 0.51
Centroid Alignment‡ (Afrasiyabi, Lalonde, and Gagn'e 2020)	*ResNet18*	74.22 ± 1.09	88.65 ± 0.55
Oblique Manifold‡ (Qi et al. 2021)	*ResNet18*	$78.24 \pm -$	$92.15 \pm -$
ECKPN♭ (Chen et al. 2021a)	*ResNet12*	77.43 ± 0.54	92.21 ± 0.41
AGAM♮ (Huang et al. 2021)	*ResNet12*	79.58 ± 0.25	87.17 ± 0.23
ADRGN♭♮ (Chen et al. 2022)	*ResNet12*	82.32 ± 0.51	92.97 ± 0.35
FCAM (ours)	*ResNet12*	$\mathbf{83.20 \pm 0.27}$	$\mathbf{93.22 \pm 0.39}$

† results are reported in (Zhang et al. 2022). ‡ methods with bigger backbones. ♯ reproduced using the data split we use. ♭ transductive methods. ♮ methods that use attribute information. The second best results are underlined.

Table 2: Comparison to the state-of-the-art methods on CUB-200-2011, ordered chronologically. Average 5-way 1-shot and 5-way 5-shot accuracy (%) with 95% confidence intervals.

FC	AM	1-shot	5-shot
		68.77 ± 0.29	84.13 ± 0.53
✓		69.40 ± 0.29	85.28 ± 0.52
	✓	69.01 ± 0.28	84.41 ± 0.53
✓	✓	$\mathbf{70.20 \pm 0.28}$	$\mathbf{85.61 \pm 0.52}$

Table 3: Ablation of Feature Calibration (FC) and Adaptive Metric (AM).

Setting	1-shot	5-shot
CKD	69.94 ± 0.28	84.79 ± 0.53
SKD	$\mathbf{70.20 \pm 0.28}$	$\mathbf{85.61 \pm 0.52}$
w/o RM	69.69 ± 0.28	85.23 ± 0.52
w/ RM	$\mathbf{70.20 \pm 0.28}$	$\mathbf{85.61 \pm 0.52}$

Table 4: Comparison of using classical KL-Divergence (CKD) and the proposed SKD for distillation (top), and the results of whether using RM to adjust ε (bottom).

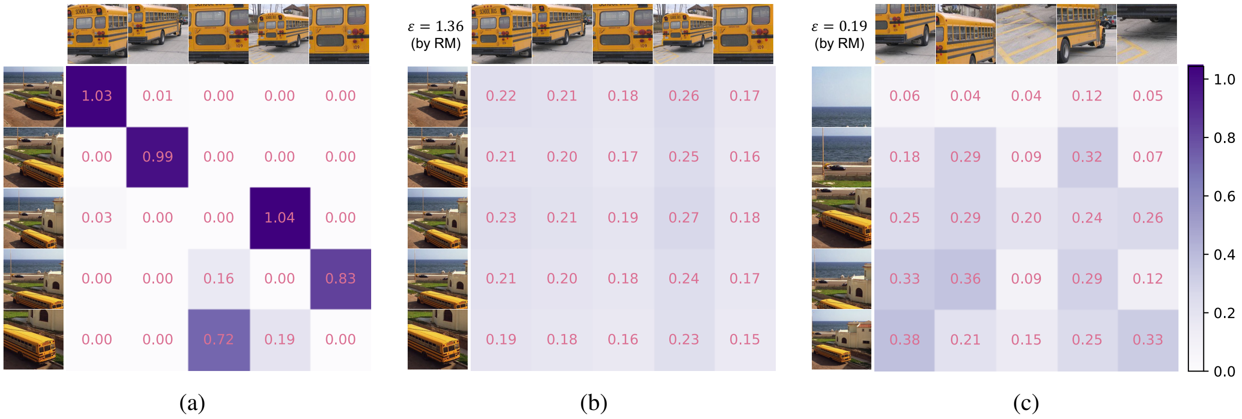

Figure 7: Visualization of solved transport matrices. Results of (a) EMD and (b) AM for sets consisting of similar local patches, and the result of (c) AM for sets consisting of dissimilar local patches. More results are presented in the supplementary material.

adjust ε is also presented in Table 4. Compared to a fixed default value, RM introduces flexibility into the metric process, helping achieve higher performance by realizing an Adaptive Metric.

Conclusions

We presented a novel FCAM method for FSL. It calibrates the biased features towards the test scenario and measures local feature sets adaptively, unleashing the power of local representations in improving novel-class generalization.

References

Afrasiyabi, A.; Lalonde, J.-F.; and Gagn'e, C. 2020. Associative Alignment for Few-shot Image Classification. In *European Conference on Computer Vision (ECCV)*, 18–35. Springer.

Ba, J.; and Caruana, R. 2014. Do Deep Nets Really Need to be Deep? In *Advances in Neural Information Processing Systems*, volume 27.

Buciluundefined, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model Compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, 535–541. ISBN 1595933395.

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust Principal Component Analysis? *J. ACM*, 58(3).

Chen, C.; Yang, X.; Xu, C.; Huang, X.; and Ma, Z. 2021a. ECKPN: Explicit Class Knowledge Propagation Network for Transductive Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6596–6605.

Chen, C.; Yang, X.; Yan, M.; and Xu, C. 2022. Attribute-Guided Dynamic Routing Graph Network for Transductive Few-Shot Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 6259–6268. ISBN 9781450392037.

Chen, Y.; Liu, Z.; Xu, H.; Darrell, T.; and Wang, X. 2021b. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9062–9071.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1126–1135.

Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born Again Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1607–1616.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; Piot, B.; kavukcuoglu, k.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, 21271–21284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A Comprehensive Overhaul of Feature Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Hiller, M.; Ma, R.; Harandi, M.; and Drummond, T. 2022. Rethinking Generalization in Few-Shot Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv e-prints*, arXiv:1503.02531.

Hu, S. X.; Li, D.; Stühmer, J.; Kim, M.; and Hospedales, T. M. 2022. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9068–9077.

Huang, S.; Zhang, M.; Kang, Y.; and Wang, D. 2021. Attributes-Guided and Pure-Visual Attention Alignment for Few-Shot Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 7840–7847.

Kim, J.; Park, S.; and Kwak, N. 2018. Paraphrasing Complex Network: Network Compression via Factor Transfer. In *Advances in Neural Information Processing Systems*, volume 31.

Li, H.; Dong, W.; Mei, X.; Ma, C.; Huang, F.; and Hu, B.-G. 2019a. LGM-Net: Learning to Generate Matching Networks for Few-Shot Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3825–3834.

Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; and Luo, J. 2019b. Revisiting Local Descriptor Based Image-To-Class Measure for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Y.; Zhang, W.; Xiang, C.; Zheng, T.; Cai, D.; and He, X. 2022. Learning To Affiliate: Mutual Centralized Learning for Few-Shot Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14411–14420.

Ma, J.; Xie, H.; Han, G.; Chang, S.-F.; Galstyan, A.; and Abd-Almageed, W. 2021. Partner-Assisted Learning for Few-Shot Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10573–10582.

Oreshkin, B.; Rodríguez López, P.; and Lacoste, A. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, volume 31.

Qi, G.; Yu, H.; Lu, Z.; and Li, S. 2021. Transductive Few-Shot Classification on the Oblique Manifold. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8412–8422.

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 *Proceedings of 6th International Conference on Learning Representations ICLR*.

Rubner, Y.; Guibas, L.; and Tomasi, C. 1997. The Earth Mover"s Distance, MultiDimensional Scaling, and Color-Based Image Retrieval. *Proceedings of the Arpa Image Understanding Workshop*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Schwartz, E.; Karlinsky, L.; Shtok, J.; Harary, S.; Marder, M.; Kumar, A.; Feris, R.; Giryes, R.; and Bronstein, A. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems*, volume 31.

Sinkhorn, R.; and Knopp, P. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2): 343–348.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Vinyals, O.; Blundell, C.; Lillicrap, T.; kavukcuoglu, k.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*, volume 29.

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

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8012–8021.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A Gift From Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X.-S. 2020. Interventional Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 33, 2734–2746.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2022. DeepEMD: Differentiable Earth Mover's Distance for Few-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–17.

Zhang, Z.; and Sabuncu, M. 2020. Self-Distillation as Instance-Specific Label Smoothing. In *Advances in Neural Information Processing Systems*, volume 33, 2184–2195.