Unleash the Power of Local Representations: Unbiased Features and Adaptive Metric for Few-Shot Learning

Shi Tang
School of Software
Tsinghua University

Chaoqun Chu
School of Software
Tsinghua University

Guiming Luo*
School of Software
Tsinghua University

Xinchen Ye
DUT-RU ISE
Dalian University of Technology

Zhiyi Xia
School of Software
Tsinghua University

Haojie Li
DUT-RU ISE
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 of novel-class generalization. To address these issues, we investigate unbiased features (UF) and an adaptive metric (AM) to propose a novel method for FSL, namely UFAM. 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, realizing an adaptive metric. 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.

1 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 [1–8], metric-based methods [1–5] 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 will be pushed far apart in the feature space [4, 9]. Recent approaches [4, 5, 9, 10] 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

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

Preprint. Under review.

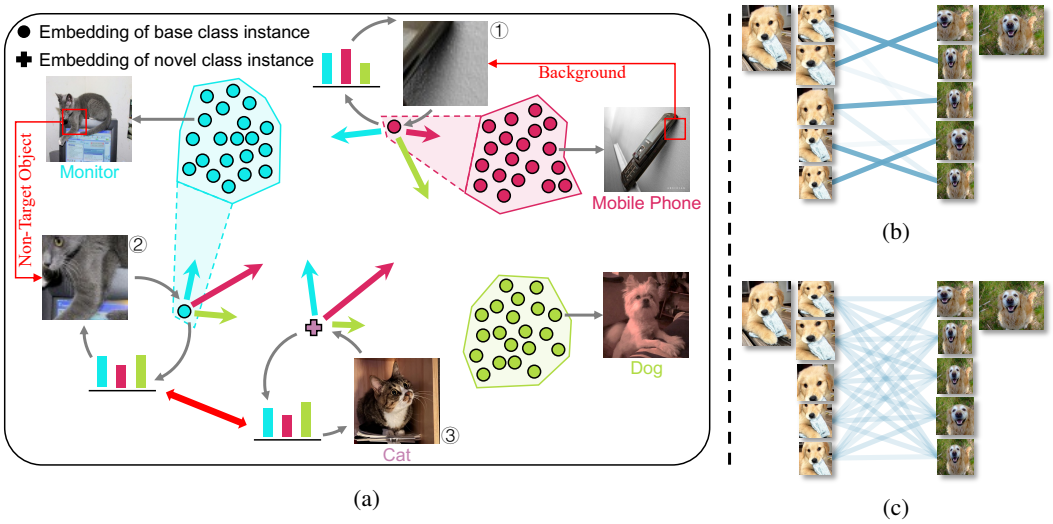


Figure 1: (a) Soft labels can be used to calibrate biased features towards the test scenario, and the visualized matching flows between two sets of similar local patches using (b) EMD and (c) dSD.

vectors [4, 5, 9, 11] or embeddings of patches cropped grid-like [4, 5] or randomly [5]. And then support-query pairs are measured by metrics capable of measuring two sets, e.g., accumulated cosine similarities between nearest neighbors [11], bidirectional random walk [4] or the Earth Mover's Distance (EMD) [5, 9].

Despite the promising results, existing methods are unable to take full advantage of 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 guide the encoder to initially acquire the ability to extract image features before episodic training [4, 5, 12, 13]. As a simple and effective augmentation, random cropping is often used in this process to increase data diversity while aligning the training objective with the encoder usage scenario, i,e., generating local features from cropped patches. However, existing methods still assume that the cropped local patches are base class samples following the traditional paradigm designed to improve intra-class generalization. This hinders the learning of high-quality local features and overlooks the potential of random cropping in improving novel-class generalization. Specifically, 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 damage generalization, which will cause performance degradation instead. Besides, due to the semantic difference, they 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 [5]. 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 [14], 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), leading to a regional monopoly. 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, by investigating unbiased features and an adaptive metric, we propose a novel method, namely UFAM, to unleash the power of local representations for few-shot 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 networks used for FSL. 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 [14] for measuring two local feature sets. It encourages smoother transport matrices capable of breaking regional monopolies as shown in Figure 1 (c), endowing the algorithm the ability to handle sets consisting of similar local features. 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:

- Unbiased features and an adaptive metric are investigated to unleash the power of local representations in improving novel-class generalization for FSL.
- A novel pretraining paradigm for FSL is proposed to calibrate the biased features towards the test scenario, and a Smoothed KL-Divergence more suitable for distilling networks for FSL is designed based on an insightful analysis of the classical KL-Divergence.
- The dual-Sinkhorn Divergence is employed as the metric to handle sets consisting of similar local features, and a Regulation Module is constructed to realize an adaptive metric.

2 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 [2, 4, 5, 9] or also meta-learned [3] metric as the classifier. For example, Snell *et al.* [2] averages the embeddings of congener support samples as the class prototype and leverages the Euclidean distance for classification. Sung *et al.* [3] 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.* [5] proposes three methods for constructing local feature sets and employs EMD for measuring two sets. Liu *et al.* [4] 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 [15–17], knowledge distillation aims at transferring "knowledge", such as logits [17] or intermediate features [18–20], 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 [21]. Zhang *et al.* [22] relate self-distillation with label smoothing, a commonly-used regularization technique to prevent models from being over-confident. Inspired by the fact that the smoothed soft labels can be used to describe "pseudo" novel class samples, we find the regularization effect of self-distillation is also beneficial in improving novel-class generalization.

Optimal transport distances. The optimal transport problem is well-studied, and the distances based on it are very powerful for probability measures. EMD was first proposed for image retrieval [23] and exhibited excellent performance. Cuturi [14] regularizes the transport 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.

3 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

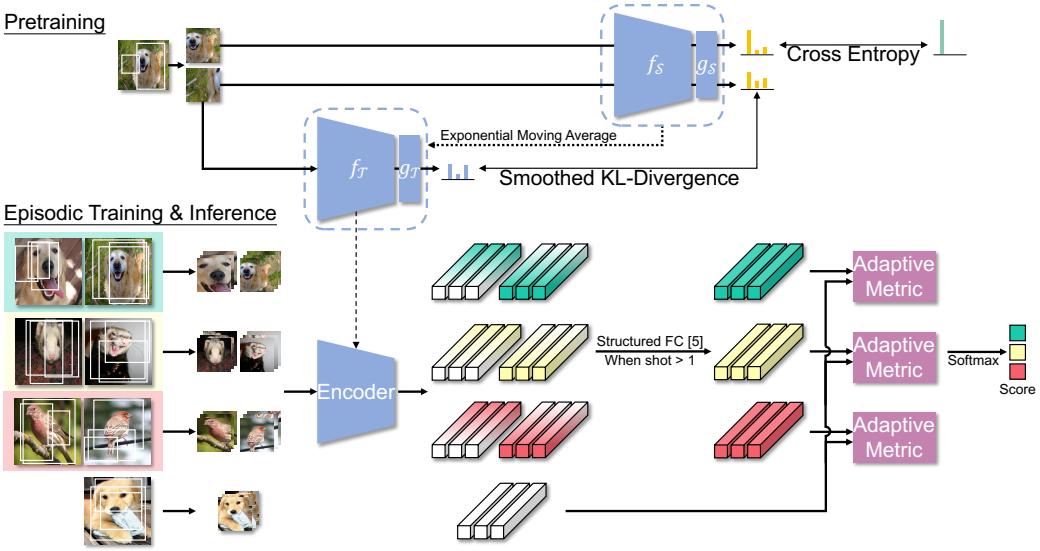


Figure 2: Overview of our framework (3-way 2-shot as an example).

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.

3.1 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.

3.2 Unbiased Features

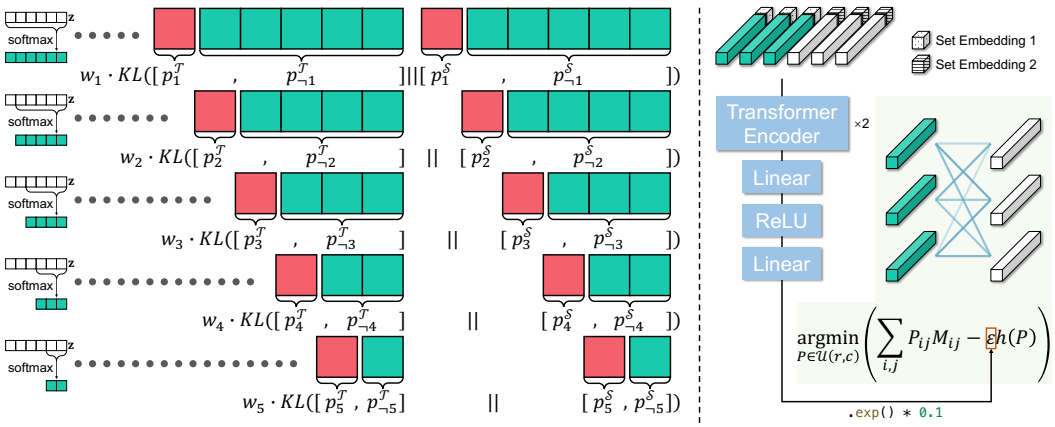
Feature calibration with self-distillation. Different from existing methods that supervise the learning of cropped local patches with ground-truth hard labels during pretraining, we advocate taking soft labels into account as well. Indicating the probability of a patch belonging to each base class, they implicitly take base class prototypes as the manifold bases to represent patches, which can properly describe the background or non-target objects[2], 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 $\phi_S = f_S \circ g_S$ and a teacher network $\phi_T = f_T \circ g_T$, with f_S and f_T being their respective encoders and g_S and g_T being their respective last FC 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 learning standard classification using the cross-entropy loss:

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

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

[2]Embeddings in the manifold space can be represented linearly or nonlinearly by the manifold bases [24, 25].



(a) (b)

Figure 3: Illustration of (a) the continuous binary classification process corresponding to the reformulation of KL-Divergence, and (b) the proposed adaptive metric based on dSD and RM.

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 [26]:

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

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 [27, 28].

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 (a), 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):

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

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 the 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 $\mathbf{b}_i$. Such a weighting scheme is not suitable for distilling a network for FSL. Because we implicitly take base class prototypes as the manifold bases to describe local patches, the probabilities of each binary classification carry important information about a manifold base and should be valued equally. 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:

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

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:

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

3.3 Adaptive Metric

After pretraining, $f_{\mathcal{T}}$ will be used as the feature extractor for further episodic training where we propose to employ dSD with RM as shown in Figure 3 (b) to produce similarity score for each support-query pair.[3]

The dual-Sinkhorn Divergence for few-shot classification. Following [5], 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 [5] followed by normalization to make both sides have the same total units for matching:

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

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

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:

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

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 $\arg\min_{P\in\mathcal{U}(\mathbf{r},\mathbf{c})} \sum_{i,j} P_{ij} M_{ij}$ [23]. Different from EMD, the dual-Sinkhorn Divergence [14] encourages a smoother transport matrix by introducing an entropic regularization:

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

[3]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 [5] and the latter process is the same as the 1-shot case.

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 [14] that can be solved with the Sinkhorn-Knopp algorithm [29] efficiently. With the solved P^{ε}, the similarity of a support-query pair can be obtained and be used to compute the classification score:

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

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 [30], considering that its inductive bias suits the task of modeling the relationship between local features very well. As shown in Figure 3 (b), 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.

4 Experiments

Datasets. The experiments are conducted on three popular benchmarks: (1) ***mini*ImageNet** [1] is a subset of ImageNet [31] 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** [32] is also a subset of ImageNet [31] 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** [33] contains 200 bird categories with 11,788 images, which represents a fine-grained scenario. Following the splits in [34], the 200 classes are divided into 100/50/50 for train/val/test respectively, and each image is first cropped with the provided human-annotated bounding box as many previous works [4, 5, 35].

Backbone. For the backbone, we employ ResNet12 as many previous works for a fair comparison. 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.

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 Adam optimizer is adopted. 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 where the SGD optimizer with a momentum of 0.9 is adopted. More detailed training settings are described in the supplementary material.

4.1 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 [4]. In particular, our method even outperforms state-of-the-art transductive [36, 37] and cross-modal [37, 38] methods, shedding some

light on how much the poor local representations can degrade the performance in the fine-grained scenario.

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	*mini*ImageNet		*tiered*ImageNet	
		1-shot	5-shot	1-shot	5-shot
MatchNet† [1]	*ResNet12*	63.08 ± 0.80	75.99 ± 0.60	68.50 ± 0.92	80.60 ± 0.71
ProtoNet† [2]	*ResNet12*	60.37 ± 0.83	78.02 ± 0.57	65.65 ± 0.92	83.40 ± 0.65
TADAM [39]	*ResNet12*	58.50 ± 0.30	76.70 ± 0.30	-	-
FEAT [34]	*ResNet12*	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
DeepEMD [9]	*ResNet12*	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
Meta-Baseline [12]	*ResNet12*	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.74 ± 0.18
FRN [10]	*ResNet12*	66.45 ± 0.19	82.83 ± 0.13	72.06 ± 0.22	86.89 ± 0.14
PAL [40]	*ResNet12*	$\underline{69.37 \pm 0.64}$	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47
MCL [4]	*ResNet12*	69.31 ± 0.20	$\underline{85.11 \pm 0.20}$	73.62 ± 0.20	86.29 ± 0.20
DeepEMD v2 [5]	*ResNet12*	68.77 ± 0.29	84.13 ± 0.53	$\underline{74.29 \pm 0.32}$	$\underline{87.08 \pm 0.60}$
Centroid Alignment‡ [41]	*WRN-28-10*	65.92 ± 0.60	82.85 ± 0.55	74.40 ± 0.68	86.61 ± 0.59
Oblique Manifold‡ [42]	*ResNet18*	63.98 ± 0.29	82.47 ± 0.44	70.50 ± 0.31	86.71 ± 0.49
FewTURE‡ [43]	*ViT-Small*	68.02 ± 0.88	84.51 ± 0.53	72.96 ± 0.92	86.43 ± 0.67
UFAM (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 [5]. ‡ methods with bigger backbones. 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.

Method	Backbone	CUB-200-2011	
		1-shot	5-shot
MatchNet† [1]	*ResNet12*	71.87 ± 0.85	85.08 ± 0.57
ProtoNet† [2]	*ResNet12*	66.09 ± 0.92	82.50 ± 0.58
DeepEMD [9]	*ResNet12*	75.65 ± 0.83	88.69 ± 0.50
FRN♯ [10]	*ResNet12*	78.86 ± 0.28	$\underline{90.48 \pm 0.16}$
MCL♯ [4]	*ResNet12*	$\underline{79.39 \pm 0.29}$	$\underline{90.48 \pm 0.49}$
DeepEMD v2 [5]	*ResNet12*	79.27 ± 0.29	89.80 ± 0.51
Centroid Alignment‡ [41]	*ResNet18*	74.22 ± 1.09	88.65 ± 0.55
Oblique Manifold‡ [42]	*ResNet18*	$78.24 \pm -$	$92.15 \pm -$
ECKPNᵖ [36]	*ResNet12*	77.43 ± 0.54	92.21 ± 0.41
AGAM♮ [38]	*ResNet12*	79.58 ± 0.25	87.17 ± 0.23
ADRGNᵖ♮ [37]	*ResNet12*	82.32 ± 0.51	92.97 ± 0.35
UFAM (ours)	*ResNet12*	$\mathbf{83.20 \pm 0.27}$	$\mathbf{93.22 \pm 0.39}$

† results are reported in [5]. ‡ methods with bigger backbones.
♯ reproduced using the data split we use. ᵖ transductive methods.
♮ methods that use attribute information. The second best results are underlined.

Table 3: Ablation of unbiased features (UF) and adaptive metric (AM). The experiments are conducted with *ResNet12* on *mini*ImageNet.

UF	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 4: Results of whether using RM to adjust ε. The experiments are conducted with *ResNet12* on *mini*ImageNet.

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

4.2 Ablation Study

We perform an ablative analysis regarding all techniques used in our method. Firstly, 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. UF adopts an additional $\mathcal{L}_{SKD}$ during pretraining, and AM employs dSD with RM for similarity measurement. With each of them outperforming the baseline and achieving optimal results when used together, their respective effectiveness can be validated. Furthermore, we conduct a more detailed analysis below.

4.2.1 Unbiased Features

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 4 (a). 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 4 (b) and (c) also illustrate that feature calibration results in better clusters for novel classes.

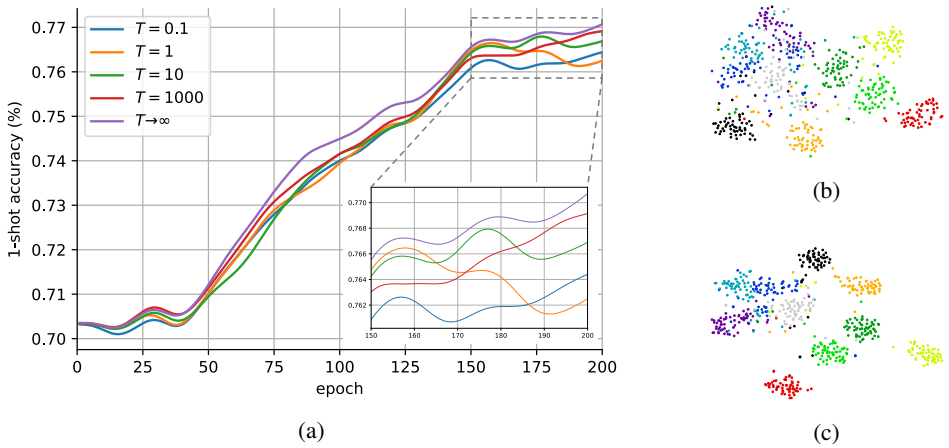


Figure 4: (a) Gaussian smoothed 1-shot test accuracy curves on CUB-200-2011 during self-distillation, with different temperatures to adjust the weighting scheme of the classical KL-Divergence. The results of the same 1000 tasks are averaged for each data point. And the t-sne visualization [44] of novel class samples embedded by encoders trained (b) with and (c) without feature calibration.

Figure 5: 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.

SKD is more suitable for feature calibration. Based on the idea of adjusting the weighting scheme of KL-Divergence by introducing a temperature coefficient, we compare different temperature settings in Figure 4 (a). 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, demonstrating the importance of a smoother weighting scheme in feature calibration.

4.2.2 Adaptive Metric

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 5 (a)) is very sparse, which

tries to match a patch with few "most" similar opposite patches. In contrast, dSD (Figure 5 (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 5 (b)). While for sets consisting of dissimilar local patches, a relatively smaller ε is predicted, making the transport matrix moderately sparse (Figure 5 (c)). Quantitative results of whether using RM to adjust ε is also presented in Table 4. Compared to a pre-fixed default value, RM introduces flexibility into the metric process, helping achieve higher performance by realizing an adaptive metric. Although such a parameterized module will inevitably bring additional time cost, since the solving of the OT problem is accelerated by dSD, the total time cost for similarity measurement does not exceed EMD by much according to a simple comparative experiment presented in the supplementary material.

5 Conclusions

In this paper, we presented a novel UFAM method for few-shot classification. It calibrates the biased features towards the test scenario and can handle various local feature sets with a designed adaptive metric. By investigating unbiased features and an adaptive metric, we managed to unleash the power of local representations to improve novel-class generalization further. Our method achieves new state-of-the-art on multiple datasets.

References

[1] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[2] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[3] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Yang Liu, Weifeng Zhang, Chao Xiang, Tu Zheng, Deng Cai, and Xiaofei He. Learning to affiliate: Mutual centralized learning for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14411–14420, June 2022.

[5] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Differentiable earth mover's distance for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2022.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 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*, pages 1126–1135, 06–11 Aug 2017.

[7] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[8] Huaiyu Li, Weiming Dong, Xing Mei, Chongyang Ma, Feiyue Huang, and Bao-Gang Hu. 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*, pages 3825–3834, 09–15 Jun 2019.

[9] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

[11] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. 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)*, June 2019.

[12] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9062–9071, October 2021.

[13] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. 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)*, pages 9068–9077, June 2022.

[14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[15] Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, 2006.

[16] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv e-prints*, page arXiv:1503.02531, March 2015.

[18] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 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)*, July 2017.

[19] Jangho Kim, Seonguk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[20] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[21] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616, 10–15 Jul 2018.

[22] Zhilu Zhang and Mert Sabuncu. Self-distillation as instance-specific label smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pages 2184–2195, 2020.

[23] Y. Rubner, L. Guibas, and C. Tomasi. The earth mover"s distance, multidimensional scaling, and color-based image retrieval. *Proceedings of the Arpa Image Understanding Workshop*, 1997.

[24] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), Jun 2011.

[25] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 2734–2746, 2020.

[26] Antti Tarvainen and Harri Valpola. 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, 2017.

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[29] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[32] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *Proceedings of 6th International Conference on Learning Representations ICLR*, 2018.

[33] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[34] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 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)*, June 2020.

[35] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[36] Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, and Zhe Ma. 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)*, pages 6596–6605, June 2021.

[37] Chaofan Chen, Xiaoshan Yang, Ming Yan, and Changsheng Xu. Attribute-guided dynamic routing graph network for transductive few-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6259–6268, 2022.

[38] Siteng Huang, Min Zhang, Yachen Kang, and Donglin Wang. Attributes-guided and pure-visual attention alignment for few-shot recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7840–7847, May 2021.

[39] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[40] Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, and Wael Abd-Almageed. Partner-assisted learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10573–10582, October 2021.

[41] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagn'e. Associative alignment for few-shot image classification. In *European Conference on Computer Vision (ECCV)*, pages 18–35. Springer, 2020.

[42] Guodong Qi, Huimin Yu, Zhaohui Lu, and Shuzhao Li. Transductive few-shot classification on the oblique manifold. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8412–8422, October 2021.

[43] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[44] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.