Unleash the Power of Local Representations for Few-Shot Classification

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*—Generalizing to novel classes unseen during training is a key challenge of few-shot classification. Recent metric-based methods try to address this by local representations. However, they are unable to take full advantage of them due to (i) biased features caused by inheriting the old paradigm of using random cropping for augmentation, and (ii) a non-adaptive metric that cannot handle various possible compositions of local feature sets. In this work, we unleash the power of local representations in improving novel-class generalization. For the encoder, we design a novel pretraining paradigm that learns cropped patches by soft labels. It avoids the semantic misalignment between hard labels and patches while fully utilizing the class-level diversity of patches. To align network output with soft labels, we also propose a Smoothed KL-Divergence that emphasizes the equal contribution of each base class in describing "non-base" patches. For the metric, we propose to predict the adjustment coefficient of an introduced entropic term to endow optimal transport distances with the necessary adaptability. Our method achieves new state-of-the-art performance on three popular benchmarks. Moreover, it exceeds state-of-the-art transductive and cross-modal methods in the fine-grained scenario, revealing how much the poor local representations can degrade the performance in this scenario.**

I. Introduction

Few-shot classification aims to distinguish between 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 congeneric samples are pushed far apart in the feature space [4], [9]. Recent approaches [4], [5], [9], [10] try to solve this 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 possible 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 [4], [5], [9], [11] or embeddings of patches cropped grid-like [4], [5] or randomly [5]. Then, support-query pairs are mesured by a metric capable of measuring two sets, e.g., accumulated cosine similarities between nearest neighbors [11], bidirectional random walk [4] or the Earch Mover's Distance (EMD) [5], [9].

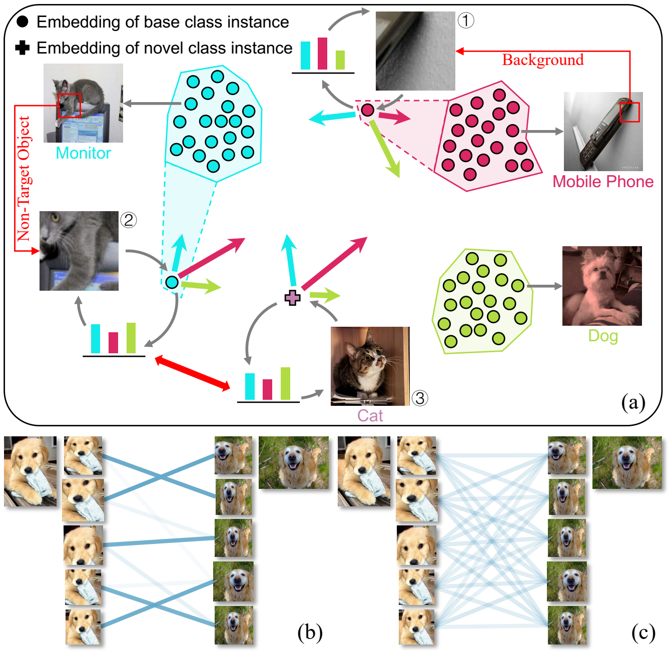
● Embedding of base class instance
✚ Embedding of novel class instance
①
Background
Monitor
Mobile Phone
Non-Target Object
②
Dog
③
Cat
(a)
(b)
(c)

Fig. 1. (a) Hard labels could provide false supervision since random cropping may alter the semantic. Describing patches by analogy, soft labels can avoid this and utilize the class-level diversity provided by random cropping. The matching flows between two sets of similar local patches using (b) EMD and (c) our Adaptive Metric.

Despite the promising results, there still exists unexploited potential of local representations in improving novel-class generalization. The first problem lies in the biased features. Typically, the encoder is first pre-trained with a proxy task that distinguishes all base classes and then meta-trained in an episodic manner [4], [5], [12], [13]. In the proxy task, random cropping is inherited as a simple and effective augmentation. However, the semantic of a patch may differ from the uncropped raw image (Fig. 1 (a) ①②). The ground-truth label can thus be false supervision, which correlats the background or non-target objects to the target object[1] and impedes learning high-quality image representations. Although this 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

[1] This is acceptable for normal intra-class classification tasks as it serves as a shortcut knowledge (e.g., dolphins are usually in the water) which improves the performance [14]. But in the few-shot setting, these priors do not hold for novel classes, which introduces bias.

samples providing class-level diversity, which can be used to prevent the network from overfitting to base classes. Moreover, non-target objects (Fig. 1 (a) ②) may be related to possible novel classes (Fig. 1 (a) ③), which can be utilized to warm up the encoder, narrowing the gap between training and testing.

The second problem lies in the metric. Better capable of dealing with the uncertainty of novel classes, random cropping also stands out in constructing local feature sets, where EMD, a well-known optimal transport (OT) distance, exhibits great superiority in measuring two sets [5]. However, it lacks the ability to handle sets consisting of similar local features, making it not adaptive enough for various possible compositions of sets caused by random cropping. Specifically, the optimum transport matrix is usually solved on a vertex of the transport polytope [15], resulting in a sparse transport matrix. As a consequence, EMD tries to match a patch with certain "most" similar opposite patch even when the opposite patches are highly similar, as shown in Fig. 1 (b). 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, as shown in Fig. 1 (c).

In this paper, by investigating Feature Calibration and Adaptive Metric, we propose a novel method, namely FCAM, to unleash the power of local representations in novel-class generalization. For Feature Calibration, we advocate using soft labels to supervise the learning of cropped patches during pretraining. Indicating the probability of a patch belonging to each base class, they are capable of describing the background or non-target objects by analogy[2] (e.g., a cat is something more like a dog and less like a monitor). This supervision makes it possible to learn from these "pseudo" novel class samples for regularization while avoiding false supervision. Moreover, soft labels connect possible novel classes with related non-target objects through similar distributions (Fig. 1 (a) ②③), making the learning of these patches a pre-search for suitable locations to embed possible novel class samples, which adapts the encoder to potential test scenario in advance. To take full advantage of soft label supervision, we also propose a Smoothed KL-Divergence. With a smoother weighting scheme paying equal attention to distinguishing each base class, it is more suitable for aligning network output and soft labels. For Adaptive Metric, we propose a Regulation Module that predicts the adjustment coefficient of an introduced entropic regularization [15] to control the smoothness of the transport matrix according to the local feature sets to be measured. 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 Adaptive Metric to unleash the power of local representations in improving novel-class generalization for few-shot classification.
- We propose a novel pretraining paradigm to calibrate the features towards the test scenario with soft labels, along with a Smoothed KL-Divergence more suitable for this supervision.
- We propose a Regulation Module to control an introduced entropic regularization for EMD, leading to a novel metric capable of handling various possible compositions of local feature sets adaptively.

II. Related Work

In this section, we give an overview of related work in three aspects, i.e., metric-based few-shot learning, self-distillation, and optimal transport distances.

A. 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] average 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] propose 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. We aim to fully utilize the unexploited potential of them in improving novel-class generalization by improving the quality of extracted local features and the adaptability of the metric.

B. Self-Distillation

First proposed for model compression [17]–[19], knowledge distillation aims at transferring "knowledge", such as logits [19] or intermediate features [20]–[22], 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 [23]. Zhang *et al.* [24] relate self-distillation with label smoothing, a commonly-used regularization technique to prevent models from being over-confident. Our idea for feature calibration requires soft labels describing the similarities of cropped patches to different base classes. Considering that the output of a well-converged classification network naturally carries this information (the probability an input belongs to each base class), we implement feature calibration in the form of self-distillation. Our results also indicate that the regularization effect of self-distillation is also beneficial for class-level generalization.

[2] The reason for this is that patch features can be represented linearly or nonlinearly by the manifold base [16] which is instantiated as mean features of base classes here.

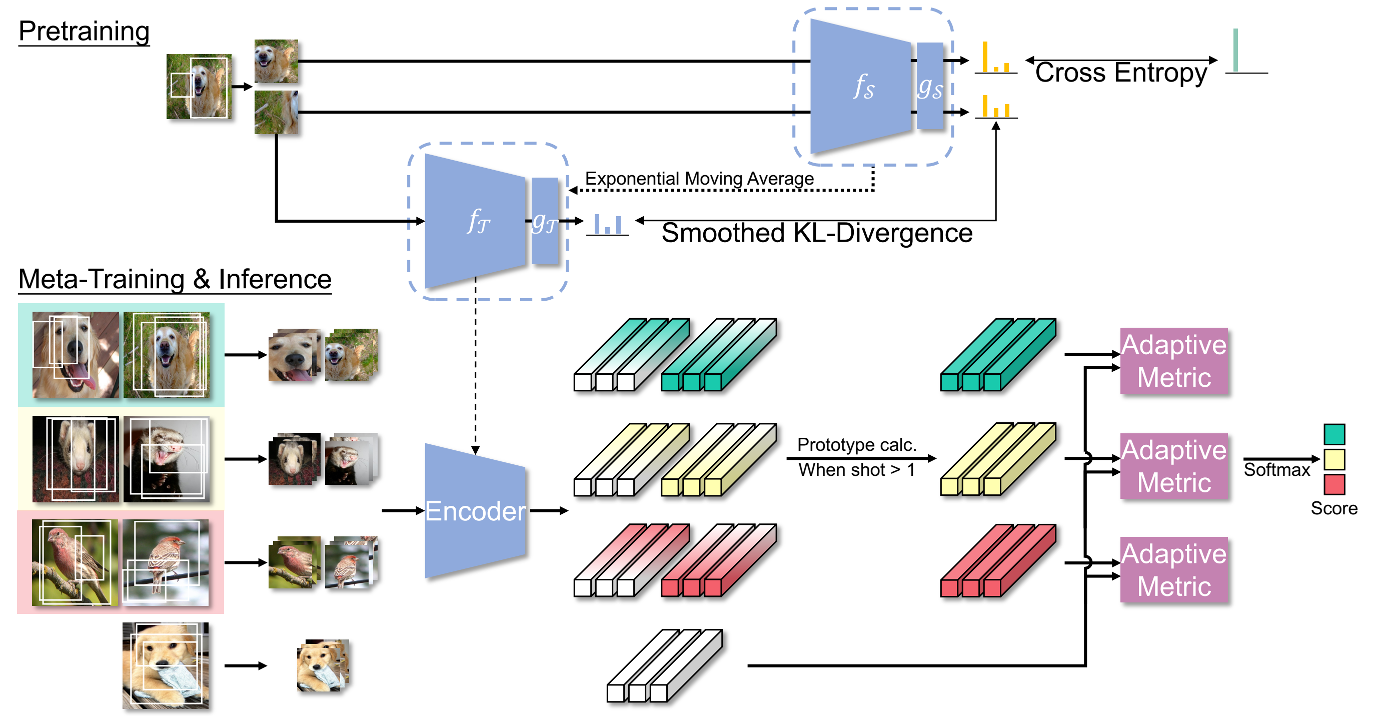

Fig. 2. Overview of our framework (3-way 2-shot as an example).

C. 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 [25] and exhibited excellent performance. Cuturi [15] proposes the dual-Sinkhorn Divergence by regularizing 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. It provides an essential prior that EMD lacks for measuring sets consisting of highly similar local features.

III. Methods

In this section, we first briefly introduce some preliminary concepts and then present our method shown in Fig. 2 in detail.

A. Preliminary

Few-shot classification aims to use a labeled dataset $D_{base} = \{(x_i^b, y_i^b)\}_{i=1}^{N_{base}}$ composed of n_c base classes to construct a classifier for future tasks consisting of novel classes. Contemporary metric-based methods usually adopt a "pretraining + meta-training" paradigm for training [4], [5], [12], [13]. In the pretraining stage, a classification network $\phi = f \circ g$ consisting of an encoder f and a linear layer g is trained to distinguish all base classes, i.e., $\phi(x_i^b) \in \mathbb{R}^{n_c}$. In the meta-training stage, the encoder is fine-tuned across a large number of N-way K-shot tasks constructed from D_{base} to simulate the test scenario. In a task containing N classes ($N < n_c$), K samples from each class are sampled to construct a support set $D_{spt} = \{(x_i^s, y_i^s)\}_{i=1}^{NK}$, according to which we need to predict labels for a query set $D_{qry} = \{(x_i^q, y_i^q)\}_{i=1}^{NQ}$ that contains samples from the same N classes with Q samples per class. Specifically, for a query sample x_i^q whose ground-truth label $y_i^q = c$ ($c \in \{1, \ldots, N\}$), the pre-trained encoder f is fine-tuned to maximize:

$$p(y_i^q = c \mid x_i^q) = \frac{\exp\left(S(R(x_i^q; f), \overline{R_c})\right)}{\sum_{j=1}^{N} \exp\left(S(R(x_i^q; f), \overline{R_j})\right)}, \qquad (1)$$

where $S(\cdot,\cdot)$ is a metric measuring the similarity between x_i^q's representation $R(x_i^q; f)$ and class j's prototype representation $\overline{R_j}$. Focusing on local representations constructed by the random copping operation $\xi(\cdot)$, for our method, $R(x_i^q; f) := \{\mathbf{u}_m | m = 1, 2, \ldots, n\}$ where $\mathbf{u}_m = f(\xi(x_i^q))$, $\overline{R_j} := \{\mathbf{v}_m | m = 1, 2, \ldots, n\}$ where $\mathbf{v}_m = \frac{1}{K}\sum_{i=1}^{NK} f(\xi(x_i^s)) \cdot [y_i^s = j]$.[3] Here, $[y_i^s = j]$ is an indicator function that equals 1 when $y_i^s = j$ and 0 otherwise.

B. Feature Calibration

1) Feature calibration with soft labels: Different from existing methods that supervise the learning of randomly cropped images with only ground-truth hard labels during pretraining, we advocate taking soft labels into account as well. As shown in Fig. 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

[3] For cases where shot $K > 1$, we conduct additional prototype calculation for a structured FC layer [5].

resize and flip). We reserve the first element $\hat{x}_1$ for normal hard label supervision using the cross-entropy loss:

$$\mathcal{L}_{CE} = -\mathbf{y}^\top \log\left(\sigma(\phi_{\mathcal{S}}(\hat{x}_1))\right), \tag{2}$$

where σ denotes the softmax function and $\mathbf{y}$ is the label of x which is a one-hot vector. The remaining $n_p - 1$ patches are used for soft label supervision, where we construct a momentum updated teacher network $\phi_{\mathcal{T}}$ to generate soft labels. Denoting 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]:

$$\theta^i_{\mathcal{T}} \leftarrow m\theta^{i-1}_{\mathcal{T}} + (1-m)\theta^i_{\mathcal{S}}, \tag{3}$$

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]. To align the output of $\phi_{\mathcal{S}}$ to that of $\phi_{\mathcal{T}}$, we propose a Smoothed KL-Divergence as described below.

2) Smoothed KL-Divergence: Denoting the network output for a patch as $\mathbf{z} = [z_1, z_2, ..., z_{n_c}] \in \mathbb{R}^{n_c}$ where z_i represents the logit of the i-th base class, the n_c-classification probabilities $\mathbf{p} = [p_1, p_2, ..., p_{n_c}] \in \mathbb{R}^{n_c}$ can be defined where the probability of the patch belonging to class i is given by:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{n_c}\exp(z_j)}. \tag{4}$$

As a common choice to measure two probability distributions, KL-Divergence is often used for soft label supervision [19], [23], [24]. However, it is not suitable for learning few-shot encoders, which we elaborate on by decomposing it.

Consider a process of continuous binary classification where each time we only focus on whether the input belongs to a certain class or to the remaining classes as illustrated in Fig. 3. The probabilities of the i-th binary classification $\mathbf{b}_i = [q_i, q_{\neg i}]$ can be obtained by:

$$q_i = \frac{\exp(z_i)}{\sum_{j=i}^{n_c}\exp(z_j)}, \quad q_{\neg i} = \frac{\sum_{k=i+1}^{n_c}\exp(z_k)}{\sum_{j=i}^{n_c}\exp(z_j)}. \tag{5}$$

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. Decomposing the multivariate distribution into a series of bivariate distributions, this decomposition helps us to investigate the probabilities for distinguishing each base class, leading to the following result.

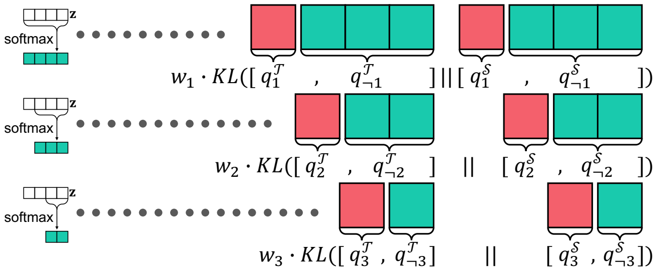
z
softmax
z
softmax
z
softmax
$w_1 \cdot KL([\,q_1^{\mathcal{T}}\ ,\ q_{\neg 1}^{\mathcal{T}}\,]\,||\,[\,q_1^{\mathcal{S}}\ ,\ q_{\neg 1}^{\mathcal{S}}\,])$
$w_2 \cdot KL([\,q_2^{\mathcal{T}}\ ,\ q_{\neg 2}^{\mathcal{T}}\,]\ ||\ [\,q_2^{\mathcal{S}}\ ,\ q_{\neg 2}^{\mathcal{S}}\,])$
$w_3 \cdot KL([\,q_3^{\mathcal{T}}\ ,\ q_{\neg 3}^{\mathcal{T}}\,]\ ||\ [\,q_3^{\mathcal{S}}\ ,\ q_{\neg 3}^{\mathcal{S}}\,])$

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

Theorem 1. *Given the respective outputs of the teacher and student, $\mathbf{z}^{\mathcal{T}}$ and $\mathbf{z}^{\mathcal{S}}$, we use the superscripts $\mathcal{T}$ and $\mathcal{S}$ to mark the variables calculated using $\mathbf{z}^{\mathcal{T}}$ and $\mathbf{z}^{\mathcal{S}}$, respectively. Then, the classical KL-Divergence for soft label supervision can be reformulated as:*

$$KL(\mathbf{p}^{\mathcal{T}}||\mathbf{p}^{\mathcal{S}}) = \sum_{i=1}^{n_c-1} w_i \cdot KL(\mathbf{b}_i^{\mathcal{T}}||\mathbf{b}_i^{\mathcal{S}}), \quad w_i = \sum_{k=i}^{n_c} p_k^{\mathcal{T}}. \tag{6}$$

Proof. According to the definition of q_i and w_i, we have:

$$q_i = \frac{\exp(z_i)}{\sum_{j=1}^{n_c}\exp(z_j)} \cdot \frac{\sum_{j=1}^{n_c}\exp(z_j)}{\sum_{j=i}^{n_c}\exp(z_j)} = \frac{p_i}{\sum_{k=i}^{n_c} p_k}, \tag{7}$$

$$p_i^{\mathcal{T}} = w_i \cdot q_i^{\mathcal{T}}. \tag{8}$$

Therefore, we have:

$$q_{\neg i} = 1 - q_i = \frac{\sum_{k=i+1}^{n_c} p_k}{\sum_{k=i}^{n_c} p_k}, \tag{9}$$

$$\sum_{k=i+1}^{n_c} p_k = q_{\neg i} \cdot \sum_{k=i}^{n_c} p_k, \tag{10}$$

$$\sum_{k=i+1}^{n_c} p_k^{\mathcal{T}} = w_i \cdot q_{\neg i}^{\mathcal{T}}. \tag{11}$$

From the above equation, it can be concluded that:

$$\sum_{k=i}^{n_c} p_k = q_{\neg(i-1)} \cdot \sum_{k=i-1}^{n_c} p_k = (\prod_{k=1}^{i-1} q_{\neg k}) \cdot (\sum_{k=1}^{n_c} p_k) = \prod_{k=1}^{i-1} q_{\neg k}. \tag{12}$$

Then, the KL-Divergence can be reformulated as follows:

$$\begin{aligned}
& KL(\mathbf{p}^{\mathcal{T}}||\mathbf{p}^{\mathcal{S}}) \\
&= \sum_{i=1}^{n_c} p_i^{\mathcal{T}} \log \frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}} \\
&= \sum_{i=1}^{n_c} p_i^{\mathcal{T}} (\log \frac{q_i^{\mathcal{T}}}{q_i^{\mathcal{S}}} + \log \frac{\sum_{k=i}^{n_c} p_k^{\mathcal{T}}}{\sum_{k=i}^{n_c} p_k^{\mathcal{S}}}) && \text{(Eq. (7))} \\
&= \sum_{i=1}^{n_c} p_i^{\mathcal{T}} \log \frac{q_i^{\mathcal{T}}}{q_i^{\mathcal{S}}} + \sum_{i=1}^{n_c} p_i^{\mathcal{T}} \log (\prod_{k=1}^{i-1} \frac{q_{\neg k}^{\mathcal{T}}}{q_{\neg k}^{\mathcal{S}}}) && \text{(Eq. (12))} \\
&= \sum_{i=1}^{n_c} w_i \cdot q_i^{\mathcal{T}} \log \frac{q_i^{\mathcal{T}}}{q_i^{\mathcal{S}}} + \sum_{i=1}^{n_c} \sum_{k=1}^{i-1} p_i^{\mathcal{T}} \log \frac{q_{\neg k}^{\mathcal{T}}}{q_{\neg k}^{\mathcal{S}}} && \text{(Eq. (8))} \\
&= \sum_{i=1}^{n_c} w_i \cdot q_i^{\mathcal{T}} \log \frac{q_i^{\mathcal{T}}}{q_i^{\mathcal{S}}} + \sum_{k=1}^{n_c-1} \sum_{i=k+1}^{n_c} p_i^{\mathcal{T}} \log \frac{q_{\neg k}^{\mathcal{T}}}{q_{\neg k}^{\mathcal{S}}} \\
&= \sum_{i=1}^{n_c} w_i \cdot q_i^{\mathcal{T}} \log \frac{q_i^{\mathcal{T}}}{q_i^{\mathcal{S}}} + \sum_{k=1}^{n_c-1} w_k \cdot q_{\neg k}^{\mathcal{T}} \log \frac{q_{\neg k}^{\mathcal{T}}}{q_{\neg k}^{\mathcal{S}}} && \text{(Eq. (11))} \\
&= \sum_{i=1}^{n_c-1} w_i \cdot (q_i^{\mathcal{T}} \log \frac{q_i^{\mathcal{T}}}{q_i^{\mathcal{S}}} + q_{\neg i}^{\mathcal{T}} \log \frac{q_{\neg i}^{\mathcal{T}}}{q_{\neg i}^{\mathcal{S}}}) \\
&= \sum_{i=1}^{n_c-1} w_i \cdot KL(\mathbf{b}_i^{\mathcal{T}}||\mathbf{b}_i^{\mathcal{S}}).
\end{aligned}$$

□

Theorem 1 demonstrates that, for KL-Divergence, the problem of measuring two probability distributions of classification can be decomposed into measuring $\mathbf{b}_i$ constantly. It also indicates how KL-Divergence weights the measurement of $\mathbf{b}_i$ by giving w_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 w_i with the class cardinality $n_c - i + 1$:

$$\tilde{w}_i = \frac{1}{n_c - i + 1} \sum_{k=i}^{n_c} p_k^{\mathcal{T}}. \tag{13}$$

According to $\tilde{w}_i$, the less similar the teacher thinks the input is to class i-n_c, the less important the alignment of $\mathbf{b}_i$. This weighting scheme is consistent with a prior of normal intra-class classification tasks, i.e., the input must belong to a certain base class. Therefore, it is reasonable to stress the measurement of $\mathbf{b}_i$ if the teacher thinks the input belongs to class i-n_c or downplay it if otherwise. However, in the context of few-shot classification, the input does not belong to any base class and measurements of different $\mathbf{b}_i$ should be equally important as each base class prototype is equal in serving as the manifold base to represent image features [29].

Noticing that the weighting scheme can be smoothed by smoothing $\mathbf{p}^{\mathcal{T}}$, we introduce a temperature coefficient T to alter the distribution of $\mathbf{p}^{\mathcal{T}}$ inspired by its use for the same purpose in various fields, e.g., contrastive learning [27] and knowledge distillation [19]. Furthermore, we discover that the difference between the weights of two different binary classifications $\tilde{w}_\alpha(T)$ and $\tilde{w}_\beta(T)$ ($\alpha \neq \beta$) vanishes with an extremely high temperature:

$$\begin{aligned} &\lim_{T\to\infty} |\tilde{w}_\alpha(T) - \tilde{w}_\beta(T)| \\ &= \lim_{T\to\infty} \left| \frac{\sum_{k_1=\alpha}^{n_c} \exp\left(z_{k_1}^{\mathcal{T}}/T\right)}{(n_c - \alpha + 1)\sum_{j=1}^{n_c} \exp\left(z_j^{\mathcal{T}}/T\right)} \right. \\ &\qquad \left. - \frac{\sum_{k_1=\beta}^{n_c} \exp\left(z_{k_2}^{\mathcal{T}}/T\right)}{(n_c - \beta + 1)\sum_{j=1}^{n_c} \exp\left(z_j^{\mathcal{T}}/T\right)} \right| = 0, \end{aligned} \tag{14}$$

according to which we derive a smoothed weighting scheme for learning few-shot encoders:

$$w_i' = \lim_{T\to\infty} \frac{\sum_{k=i}^{n_c} \exp\left(z_k^{\mathcal{T}}/T\right)}{\sum_{j=1}^{n_c} \exp\left(z_j^{\mathcal{T}}/T\right)} = \frac{n_c - i + 1}{n_c}. \tag{15}$$

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

$$SKD(\mathbf{p}^{\mathcal{T}}||\mathbf{p}^{\mathcal{S}}) := \sum_{i=1}^{n_c-1} w_i' \cdot KL(\mathbf{b}_i^{\mathcal{T}}||\mathbf{b}_i^{\mathcal{S}}), \tag{16}$$

$$\mathcal{L}_{SKD} = \frac{1}{n_p - 1} \sum_{i=2}^{n_p} SKD(\sigma(\phi_T(\hat{x}_i))||\sigma(\phi_S(\hat{x}_i))). \tag{17}$$

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

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{SKD}. \tag{18}$$

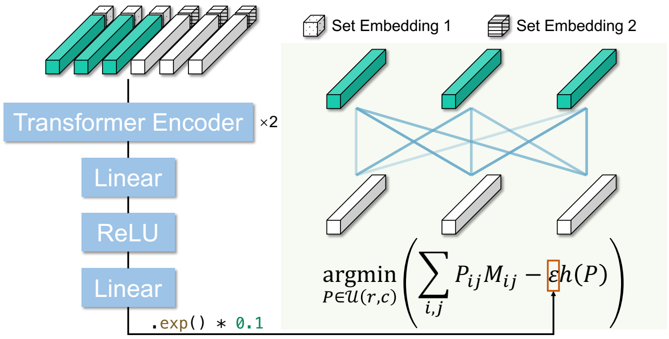
Set Embedding 1
Set Embedding 2
Transformer Encoder
×2
Linear
ReLU
Linear
.exp() * 0.1
$\underset{P\in\mathcal{U}(r,c)}{\text{argmin}}\left(\sum_{i,j} P_{ij}M_{ij} - \varepsilon h(P)\right)$

Fig. 4. The proposed Adaptive Metric formulate the measurement process as an OT problem. To handle various possible compositions of local feature sets, the adjustment coefficient of an entropy regularization is tuned by a Regulation Module.

C. Adaptive Metric

After pretraining, $f_{\mathcal{T}}$ will be used as the feature extractor for further meta-training where we propose an adaptive metric as shown in Fig. 4.

To measure $R(x_i^q; f_{\mathcal{T}})$ and $\overline{R_j}$ defined in Sec. III-A, we formulate the metric as an OT problem that considers a hypothetical process of transporting goods from nodes of one set (suppliers) to nodes of the other set (demanders). Corresponding to the importance of each local feature in a set, the total supply (demand) units of the i-th supplier r_i (demander c_i) are given by the cross-reference mechanism [5] followed by normalization to make both sides have the same total units for matching:

$$\hat{r}_i = \max\left\{\frac{1}{n}\sum_{j=1}^{n} \mathbf{u}_i^\top \mathbf{v}_j, 0\right\}, \quad r_i = \frac{n\hat{r}_i}{\sum_{j=1}^{n} \hat{r}_j}, \tag{19}$$

$$\hat{c}_i = \max\left\{\frac{1}{n}\sum_{j=1}^{n} \mathbf{v}_i^\top \mathbf{u}_j, 0\right\}, \quad c_i = \frac{n\hat{c}_i}{\sum_{j=1}^{n} \hat{c}_j}. \tag{20}$$

Given the cost to transport a unit from node u_i to v_j:

$$M_{ij} = 1 - \frac{\mathbf{u}_i^\top \mathbf{v}_j}{\|\mathbf{u}_i\|\|\mathbf{v}_j\|}, \tag{21}$$

the goal of the OT problem is to find a transportation plan with the lowest total cost from a set of valid plans $\mathcal{U}(\mathbf{r}, \mathbf{c}) = \{P \in \mathbb{R}_+^{n\times n} | P\mathbf{1}_n = \mathbf{r}, P^\top \mathbf{1}_n = \mathbf{c}\}$.

Instead of solving $\arg\min_{P\in\mathcal{U}(\mathbf{r},\mathbf{c})} \sum_{i,j} P_{ij}M_{ij}$ directly like EMD [25], we introduce an entropic regularization proposed by Cuturi [15] to encourage smoother transport matrices:

$$P^{\varepsilon} = \underset{P\in\mathcal{U}(\mathbf{r},\mathbf{c})}{\arg\min}\left(\sum_{i,j} P_{ij}M_{ij} - \varepsilon h(P)\right), \tag{22}$$

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 solved matrix. This regularization not only endows the metric with the ability to handle sets consisting of similar local features but also makes it possible to solve the transport problem faster as it becomes a strictly convex problem [15] that can be solved with the Sinkhorn-Knopp algorithm [30] efficiently.

Different from [15] that treats ε as a pre-fixed hyperparameter, we design a Regulation Module to predict it. By giving higher ε, P^{ε} will be smoother, and as ε goes to zero, it will be sparser, with the solution close to EMD. This way, we can control the smoothness of the transport matrix adaptively according to specific 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 extracted local features as input and construct a predictor based on the Transformer encoder [31], considering that its inductive bias suits the task of modeling the relationship between local features very well. As shown in Fig. 4, the input embeddings are constructed by concatenating the local feature with a 16 dimensional learnable set embedding indicating which set the local feature is from, i.e., $R(x_i^q; f_{\mathcal{T}})$ or $\overline{R_j}$. Followed by an exponential function, the output serves as a scaling factor to adjust ε based on the default value of 0.1.

Eventually, with the solved P^{ε}, we define Adaptive Metric to compute the classification score:

$$S(R(x_i^q; f_{\mathcal{T}}), \overline{R_j}) \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{n}(1 - M_{ij})P_{ij}^{\varepsilon}. \tag{23}$$

With an additional softmax function, the classification probabilities are obtained for cross entropy calculation or inference. The overall training process of our method is described in Algorithm 1.

Algorithm 1 Training process of FCAM.

```
PRETRAINING
Train $\phi_{\mathcal{S}}$ with $\mathcal{L}_{CE}$ for warmup-epochs;
$\theta_{\mathcal{T}} \leftarrow \theta_{\mathcal{S}}$;
while epochs do
  while steps do
    Randomly crop $n_p$ patches $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{n_p}\}$ for each image $x$ in the minibatch;
    Calculate $\mathcal{L}_{CE}$ with $\hat{x}_1$;
    Calculate $\mathcal{L}_{SKD}$ with $\{\hat{x}_2, \hat{x}_3, \ldots, \hat{x}_{n_p}\}$;
    $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{SKD}$;
    Update $\theta_{\mathcal{S}}$ with $\nabla_{\theta_{\mathcal{S}}}\mathcal{L}_{total}$;
    Update $\theta_{\mathcal{T}}$, i.e., $\theta_{\mathcal{T}}^{i} \leftarrow m\theta_{\mathcal{T}}^{i-1} + (1-m)\theta_{\mathcal{S}}^{i}$;
  end while
end while
META-TRAINING
while not converged do
  Construct a task, i.e., sample $D_{spt}$, $D_{qry}$ from $D_{base}$;
  Calculate $\overline{R_j}$ for $j$ in $N$;
  for $x_i^q$ in $D_{qry}$ do
    Calculate $R(x_i^q; f_{\mathcal{T}})$;
    Predict $\varepsilon$ and calculate $S(R(x_i^q; f_{\mathcal{T}}), \overline{R_j})$ for $j$ in $N$;
  end for
  Calculate cross entropy loss;
  Optimize $f_{\mathcal{T}}$;
end while
```

IV. Experiments

Datasets. The experiments are conducted on three popular benchmarks: (1) ***mini*ImageNet** [1] is a subset of ImageNet [32] 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** [33] is also a subset of ImageNet [32] 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** [34] contains 200 bird categories with 11,788 images, which represents a fine-grained scenario. Following the splits in [35], 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 Regulation Module.

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 meta-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. The Regulation Module is first 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.

A. 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 I. 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 II. 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.

For the cross-domain setting which poses a greater challenge for novel-class generalization, we perform an experiment where models are trained on *mini*Imagenet and evaluated on CUB-200-2011 following the setups in [39]. The cross-domain setting allows us to better evaluate the model's ability to handle novel classes with significant domain differences from the base classes, due to the large domain gap. As a result, it better reflects novel-class generalization. As shown in Table III, our method outperforms the previous state-of-the-art approach, demonstrating the superiority of our method in improving novel-class generalization.

TABLE I
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 [40]	*ResNet12*	58.50 ± 0.30	76.70 ± 0.30	-	-
FEAT [35]	*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 [41]	*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‡ [42]	*WRN-28-10*	65.92 ± 0.60	82.85 ± 0.55	74.40 ± 0.68	86.61 ± 0.59
Oblique Manifold‡ [43]	*ResNet18*	63.98 ± 0.29	82.47 ± 0.44	70.50 ± 0.31	86.71 ± 0.49
FewTURE‡ [44]	*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 [5]. ‡ methods with bigger backbones. The second best results are underlined.

TABLE II
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$^\sharp$ [10]	*ResNet12*	78.86 ± 0.28	$\underline{90.48 \pm 0.16}$
MCL$^\sharp$ [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‡ [42]	*ResNet18*	74.22 ± 1.09	88.65 ± 0.55
Oblique Manifold‡ [43]	*ResNet18*	$78.24 \pm -$	$92.15 \pm -$
ECKPN$^\flat$ [36]	*ResNet12*	77.43 ± 0.54	92.21 ± 0.41
AGAM$^\natural$ [38]	*ResNet12*	79.58 ± 0.25	87.17 ± 0.23
ADRGN$^{\flat\natural}$ [37]	*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 [5]. ‡ methods with bigger backbones.
$^\sharp$ reproduced using the data split we use. $^\flat$ transductive methods.
$^\natural$ methods that use attribute information.
The second best results are underlined.

TABLE III
CROSS-DOMAIN EXPERIMENTS (*mini*IMAGENET→CUB) FOLLOWING THE SETTING OF [39]. AVERAGE 5-WAY 1-SHOT AND 5-WAY 5-SHOT ACCURACY (%) WITH 95% CONFIDENCE INTERVALS.

Method	1-shot	5-shot
ProtoNet† [2]	50.01 ± 0.82	72.02 ± 0.67
MatchNet† [1]	51.65 ± 0.84	69.14 ± 0.72
cosine classifier [39]	44.17 ± 0.78	69.01 ± 0.74
linear classifier [39]	50.37 ± 0.79	73.30 ± 0.69
KNN [11]	50.84 ± 0.81	71.25 ± 0.69
DeepEMD v2 [5]	$\underline{54.24 \pm 0.86}$	$\underline{78.86 \pm 0.65}$
FCAM (ours)	$\mathbf{58.20 \pm 0.30}$	$\mathbf{80.92 \pm 0.65}$

† results are reported in [5]. The second best results are underlined.

TABLE IV
ABLATION OF FEATURE CALIBRATION AND ADAPTIVE METRIC.

Feature Calibration	Adaptive Metric	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
✓	✓	$\mathbf{70.20 \pm 0.28}$	$\mathbf{85.61 \pm 0.52}$

B. Ablation Study

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

1) Feature Calibration improves novel-class generalization: To demonstrate that Feature Calibration improves novel-class generalization, we visualize the 1-shot test accuracy change during feature calibration in Fig. 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 calibration. 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 Fig. 6 also illustrate that Feature Calibration results in better clusters for novel classes.

2) Smoothed KL-Divergence is more suitable for Feature Calibration: We compare different temperature settings in Fig. 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

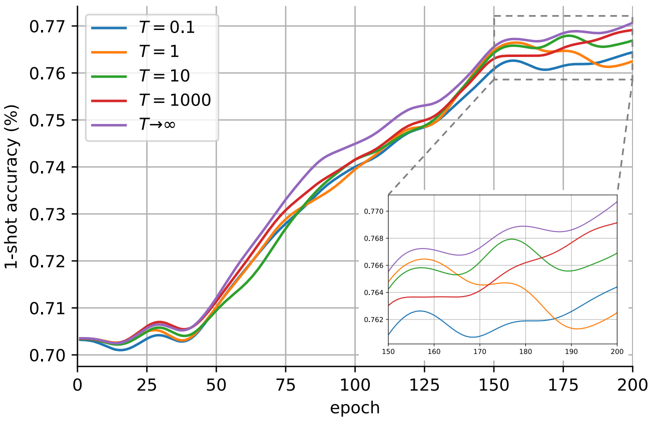

Fig. 5. Gaussian smoothed 1-shot test accuracy curves on CUB-200-2011 during feature calibration, with different temperatures to adjust the weighting scheme of the classical KL-Divergence. The results of the same 1000 tasks are averaged for each data point.

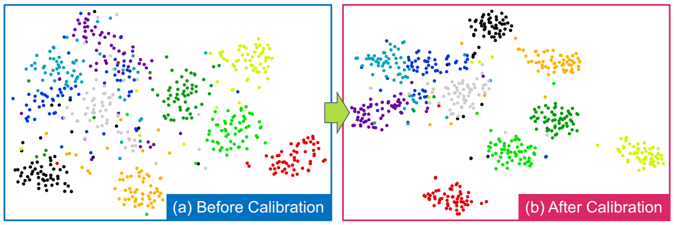

Fig. 6. The t-sne visualization [45] of novel class samples embedded by encoders trained (a) without and (b) with Feature Calibration.

Smoothed KL-Divergence, i.e., $T \to \infty$, constantly outperforms other settings. Furthermore, Smoothed KL-Divergence yields better final performance than classical KL-Divergence as shown in Table V. Both the above experiments demonstrate the importance of a smoother weighting scheme in Feature Calibration.

TABLE V
COMPARISON OF USING CLASSICAL KL-DIVERGENCE AND SMOOTHED KL-DIVERGENCE FOR DISTILLATION (TOP), AND THE RESULTS OF WHETHER USING REGULATION MODULE TO ADJUST ε (BOTTOM).

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

3) Entropic regularization handles sets consisting of similar local features: For sets consisting of similar local features, the transport matrix solved by EMD (Fig. 7 (a)) is very sparse, which tries to match a feature with few "most" similar opposite features. In contrast, our metric (Fig. 7 (b)) is able to generate a smoother transport matrix due to the entropic regularization, which enables a comprehensive utilization of opposite features and reduces the dependency on a few opposite features by allowing "one-to-many" matching.

C. Regulation Module brings adaptability

For sets consisting of similar local features, Regulation Module predicts a relatively larger ε, resulting in a smoother transport matrix (Fig. 7 (b)). While for sets consisting of dissimilar local features, a relatively smaller ε is produced, making the transport matrix moderately sparse (Fig. 7 (c)). Quantitative results of whether using Regulation Module to adjust ε is also presented in Table V. Compared to a fixed default value, it introduces flexibility into the metric process, helping achieve higher performance by realizing an adaptive measurement.

D. Analysis on computational time

Although Regulation Module will inevitably bring additional time overhead during inference as a parameterized module, the proposed Adaptive Metric still costs less time for measuring two feature sets compared to EMD since the solving of the OT problem is accelerated. Specifically, the introduced entropic regularization makes the OT problem a strictly convex problem [15]. Thus, it can be solved by the Sinkhorn-Knopp algorithm [30] which is known to have a linear convergence [46], [47]. We conduct an experiment on an RTX-3090 (Linux, PyTorch 3.6) using the same $10,000$ randomly sampled episodes to compare the time cost empirically. The average time spent to process an episode is reported in Table VI. It can be seen that our Adaptive Metric spends way less time than EMD (26.33% faster) even in the presence of a parameterized module, demonstrating its superiority in both accuracy and speed.

TABLE VI
TIME SPENT PROCESSING AN EPISODE FOR METHODS WITH DIFFERENT METRICS. 9 PATCHES ARE USED TO REPRESENT A SAMPLE.

Metric	Average time per episode (ms)
EMD	378.42
Adaptive Metric (ours)	**278.78**

V. CONCLUSIONS

In this paper, we presented a novel FCAM method for few-shot classification. It calibrates the features towards the test scenario and can handle various possible compositions of local feature sets with a proposed Adaptive Metric. By investigating Feature Calibration 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] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf

[2] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf

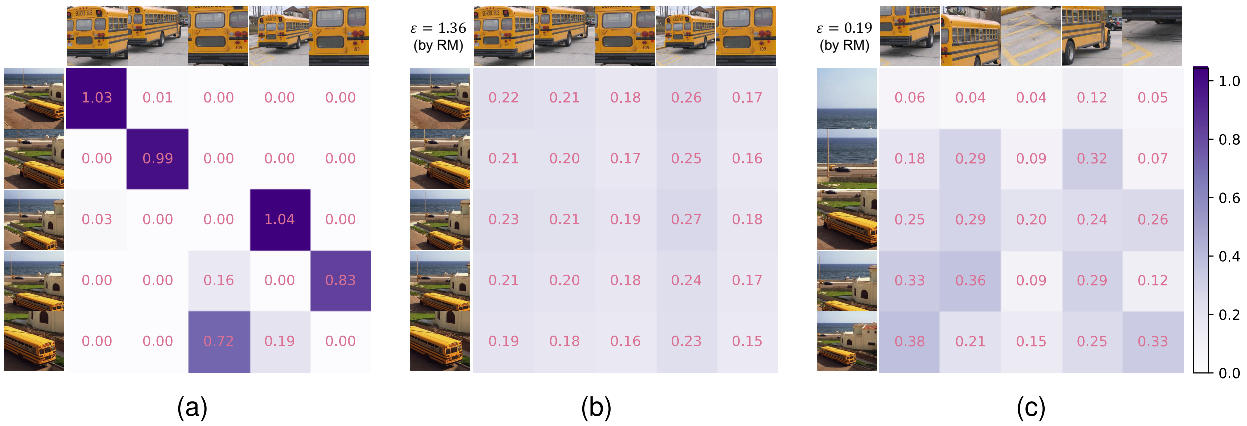

Fig. 7. Visualization of solved transport matrices. Results of (a) EMD and (b) Adaptive Metric for sets consisting of similar local features, and the result of (c) Adaptive Metric for sets consisting of dissimilar local features. More results are presented in the supplementary material.

[3] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. 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] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, and X. 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)*, June 2022, pp. 14 411–14 420.

[5] C. Zhang, Y. Cai, G. Lin, and C. Shen, “Deepemd: Differentiable earth mover’s distance for few-shot learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2022.

[6] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70, 06–11 Aug 2017, pp. 1126–1135. [Online]. Available: https://proceedings.mlr.press/v70/finn17a.html

[7] E. Schwartz, L. Karlinsky, J. Shtok, S. Harary, M. Marder, A. Kumar, R. Feris, R. Giryes, and A. Bronstein, “Delta-encoder: an effective sample synthesis method for few-shot object recognition,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/1714726c817af50457d810aae9d27a2e-Paper.pdf

[8] H. Li, W. Dong, X. Mei, C. Ma, F. Huang, and B.-G. Hu, “LGM-net: Learning to generate matching networks for few-shot learning,” in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97, 09–15 Jun 2019, pp. 3825–3834. [Online]. Available: https://proceedings.mlr.press/v97/li19c.html

[9] C. Zhang, Y. Cai, G. Lin, and C. 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] D. Wertheimer, L. Tang, and B. Hariharan, “Few-shot classification with feature map reconstruction networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8012–8021.

[11] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. 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] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, “Meta-baseline: Exploring simple meta-learning for few-shot learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9062–9071.

[13] S. X. Hu, D. Li, J. Stühmer, M. Kim, and T. 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)*, June 2022, pp. 9068–9077.

[14] K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry, “Noise or signal: The role of image backgrounds in object recognition,” in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gl3D-xY7wLq

[15] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Advances in Neural Information Processing Systems*, vol. 26, 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf

[16] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *J. ACM*, vol. 58, no. 3, Jun 2011. [Online]. Available: https://doi.org/10.1145/1970392.1970395

[17] C. Buciluundefined, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’06, 2006, p. 535–541. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[18] J. Ba and R. Caruana, “Do deep nets really need to be deep?” in *Advances in Neural Information Processing Systems*, vol. 27, 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf

[19] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” *arXiv e-prints*, p. arXiv:1503.02531, Mar. 2015.

[20] J. Yim, D. Joo, J. Bae, and J. 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.

[21] J. Kim, S. Park, and N. Kwak, “Paraphrasing complex network: Network compression via factor transfer,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/6d9cb7de5e8ac30bd5e8734bc96a35c1-Paper.pdf

[22] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, “A comprehensive overhaul of feature distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[23] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, “Born again neural networks,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, 10–15 Jul 2018, pp. 1607–1616. [Online]. Available: https://proceedings.mlr.press/v80/furlanello18a.html

[24] Z. Zhang and M. Sabuncu, “Self-distillation as instance-specific label smoothing,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 2184–2195. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1731592aca5fb4d789c4119c65c10b4b-Paper.pdf

[25] 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.

[26] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf

[27] K. He, H. Fan, Y. Wu, S. Xie, and R. 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] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot,

k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21 271–21 284. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf

[29] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, "Interventional few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 2734–2746. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1cc8a8ea51cd0adddf5dab504a285915-Paper.pdf

[30] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, vol. 21, no. 2, pp. 343–348, 1967.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

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

[33] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *Proceedings of 6th International Conference on Learning Representations ICLR*, 2018.

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[35] H.-J. Ye, H. Hu, D.-C. Zhan, and F. 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.

[36] C. Chen, X. Yang, C. Xu, X. Huang, and Z. 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)*, June 2021, pp. 6596–6605.

[37] C. Chen, X. Yang, M. Yan, and C. Xu, "Attribute-guided dynamic routing graph network for transductive few-shot learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22, 2022, p. 6259–6268. [Online]. Available: https://doi.org/10.1145/3503161.3548301

[38] S. Huang, M. Zhang, Y. Kang, and D. Wang, "Attributes-guided and pure-visual attention alignment for few-shot recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, pp. 7840–7847, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16957

[39] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *International Conference on Learning Representations*, 2019.

[40] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/66808e327dc79d135ba18e051673d906-Paper.pdf

[41] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, and W. Abd-Almageed, "Partner-assisted learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 573–10 582.

[42] A. Afrasiyabi, J.-F. Lalonde, and C. Gagn'e, "Associative alignment for few-shot image classification," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 18–35.

[43] G. Qi, H. Yu, Z. Lu, and S. Li, "Transductive few-shot classification on the oblique manifold," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8412–8422.

[44] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [Online]. Available: https://openreview.net/forum?id=p_g2nHlMus

[45] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[46] J. Franklin and J. Lorenz, "On the scaling of multidimensional matrices," *Linear Algebra and its Applications*, vol. 114-115, pp. 717–735, 1989. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0024379589904904

[47] P. A. Knight, "The sinkhorn–knopp algorithm: Convergence and applications," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 1, pp. 261–275, 2008. [Online]. Available: https://doi.org/10.1137/060659624