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], and Haojie Li[2]

[1] School of Software, Tsinghua University, Beijing
[2] DUT-RU ISE, Dalian University of Technology, Dalian

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) improper supervision for pretraining the feature extractor, and (ii) lack of adaptability in the metric for handling various possible compositions of local feature sets. In this work, we unleash the power of local representations in improving novel-class generalization. For the feature extractor, we design a novel pretraining paradigm that learns randomly cropped patches by soft labels. It utilizes the class-level diversity of patches while diminishing the impact of their semantic misalignments to hard labels. To align network output with soft labels, we also propose a UniCon KL-Divergence that emphasizes the equal contribution of each base class in describing “non-base” patches. For the metric, we formulate measuring local feature sets as an entropy-regularized optimal transport problem to introduce the ability to handle sets consisting of homogeneous elements. Furthermore, we design a Modulate Module to endow the metric 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.

Keywords: Few-shot classification · Metric learning · Optimal transport distance

1 Introduction

Given abundant samples of some classes (often called base classes) for training, few-shot classification (FSC) aims at constructing a classifier to distinguish between novel classes unseen during training with limited examples. Suffering from the low-data regimes and the inconsistency between training with base classes and testing with novel classes, FSC algorithms often struggle with poor generalization to novel classes [22, 44]. Recent approaches [22, 39, 44, 45] 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.

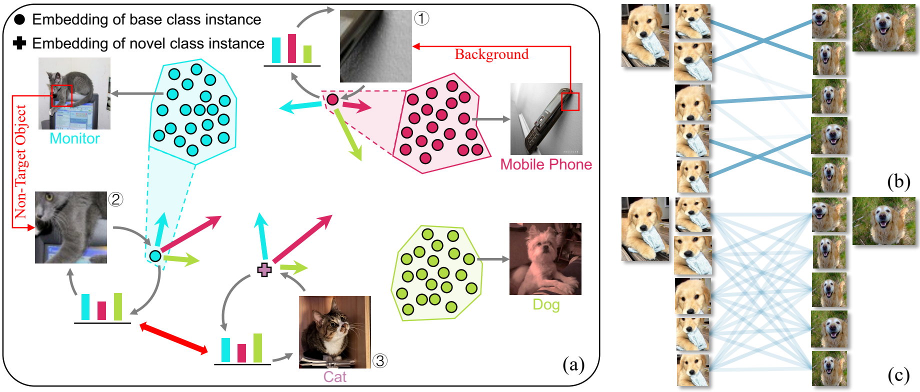
● Embedding of base class instance
✚ Embedding of novel class instance
①
Background
Monitor
Mobile Phone
Non-Target Object
②
Dog
Cat
③
(a)
(b)
(c)

Fig. 1: (a) Hard labels could provide false supervision since random cropping may alter the semantics. 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.

Within the metric learning framework, this genre relies heavily on the quality of local features and the set metric. However, both aspects are not satisfactory in existing methods, resulting in an unexploited potential of local representations in improving novel-class generalization.

Hard labels are insufficient for pretraining. Usually, the encoder for feature extraction will be pre-trained with a proxy task classifying all base classes, where random cropping is often inherited as a simple and effective augmentation [8,17,22,45]. However, it may alter the semantics (Fig. 1 (a) ①②), making ground-truth hard labels insufficient for pretraining. The reasons are, firstly, they may provide false supervision that correlates the background or non-target objects to a base class. 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 [40]. But in the few-shot setting, these priors do not hold for novel classes, which introduces bias. Secondly, they cannot utilize the class-level diversity provided by random copping to prevent the network from overfitting to base classes. Because hard labels strictly assume that the input belongs to one of the base classes, which cannot describe patches with semantics beyond all base classes.

Indicating the probability of the input belonging to each base class, soft labels are capable of describing cropped patches by analogy[3] (*e.g.*, a cat is something more like a dog and less like a monitor). Therefore, they can be used to supervise the learning of these cropped patches for regularization while avoiding false supervision. Moreover, soft labels connect non-target objects with potential novel classes through similar distributions (Fig. 1 (a) ②③), making the learning of

[3] The reason for this is that patch features can be represented linearly or nonlinearly by the manifold base [4] which is instantiated as mean features of base classes here.

these patches a pre-search for suitable positions to embed novel class samples, which warms up the encoder for possible test scenarios in advance.

Metric necessitates adaptability for various set composition. Optimal transport (OT) distances are a family of well-studied set metrics and the Earth Mover's Distance (EMD), a classic representative of them, exhibits great superiority in measuring local feature sets [44]. Recently, embedding randomly cropped patches stands out in constructing local feature sets [45] due to its ability to handle the uncertainty of novel classes with randomness. However, random cropping unavoidably results in various set compositions, *i.e.*, the local features in a set may be highly similar or completely different. But EMD lacks the ability to handle sets consisting of similar features, as it tends to produce sparse transport matrices, trying to match a patch with a certain "most" similar one in the other set, even when its patches are nearly identical (Fig. 1 (b)). A smoother transport matrix that allows "one-to-many" matching (Fig. 1 (c)) is desired under such circumstances because it reduces the dependency on a particular opposite patch, mitigating the impact of its specific cluttered background by utilizing opposite patches comprehensively.

Introducing an entropic regularization, the Sinkhorn Distance [9] encourages smoother transport matrices when solving an OT problem. Therefore, formulating the metric as an entropy-regularized OT problem can endow the algorithm with the ability to handle sets consisting of similar features. Moreover, it is possible to control the sparseness of the transport matrix by adjusting the regularization strength accordingly, which can introduce adaptability into the metric to cope with various set compositions.

In this paper, we propose a novel method for few-shot classification to unleash the power of local representations in improving novel-class generalization. To obtain better local features, we propose Feature Calibration to pre-train few-shot encoders. It supervises the learning of cropped patches with soft labels produced by a momentum-updated teacher, utilizing the class-level diversity of them while avoiding false supervision. In addition, by decomposing the classical KL-Divergence commonly used for soft label supervision, we find its inherent weighting scheme unsuitable for learning few-shot encoders as it implicitly assumes that the input must belong to a certain base class. Therefore, we propose UniCon KL-Divergence (UKD) with a more suitable weighting scheme for the soft label supervision. To measure local feature sets with various compositions, we propose Adaptive Metric that formulates the set measurement problem as a regularized OT problem, where a Modulate Module is designed to adjust the regularization strength adaptively. The proposed method, namely FCAM, 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. Our contributions are as follows:

- We propose a novel pretraining paradigm for few-shot encoders that uses soft labels to utilize the class-level diversity provided by random cropping while avoiding improper supervision.

- We propose a UniCon KL-Divergence for the soft label supervision to correct an assumption of conventional KL-Divergence that does not hold true for the few-shot setting.
- We propose a novel metric capable of handling various compositions of local feature sets adaptively for local-representation-based FSC.

2 Related Work

Metric-based few-shot classification. The literature exhibits significant diversity in the area of FSC [10,20,22,30,33,34,37,45], among which metric-based methods [22, 33, 34, 37, 45] are very elegant and promising. The main idea is to meta-learn a representation expected to be generalizable across categories with a predefined [22,33,37,44,45] or also meta-learned [34] metric as the classifier. For example, Snell *et al.* [33] average the embeddings of congener support samples as the class prototype and leverages the Euclidean distance for classification. Sung *et al.* [34] 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 resort to local representations. Generally, an instance is represented by a local feature set whose elements can be implemented as local feature vectors [21, 22, 44, 45] or embeddings of patches cropped grid-like [22,45] or randomly [45]. Then, support-query pairs are measured by a metric capable of measuring two sets, *e.g.*, accumulated cosine similarities between nearest neighbors [21], bidirectional random walk [22] or EMD [44,45].

Self-distillation. First proposed for model compression [2,3,16], knowledge distillation aims at transferring "knowledge", such as logits [16] or intermediate features [14, 19, 42], 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 [11]. Zhang *et al.* [46] relate self-distillation with label smoothing, a commonly-used regularization technique to prevent models from being overconfident. Generating soft labels with a momentum-updated teacher, our Feature Calibration is closely related to self-distillation and exhibits a similar regularization effect for improving class-level generalization.

Optimal transport distances. The distances based on the well-studied OT problem are very powerful for probability measures. EMD was first proposed for image retrieval [28] and exhibited excellent performance. Cuturi [9] proposes the Sinkhorn Distance 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.

3 Background

As illustrated in Fig. 2, we integrate our method into the "pretraining + meta-training" paradigm commonly used by contemporary metric-based methods [8,

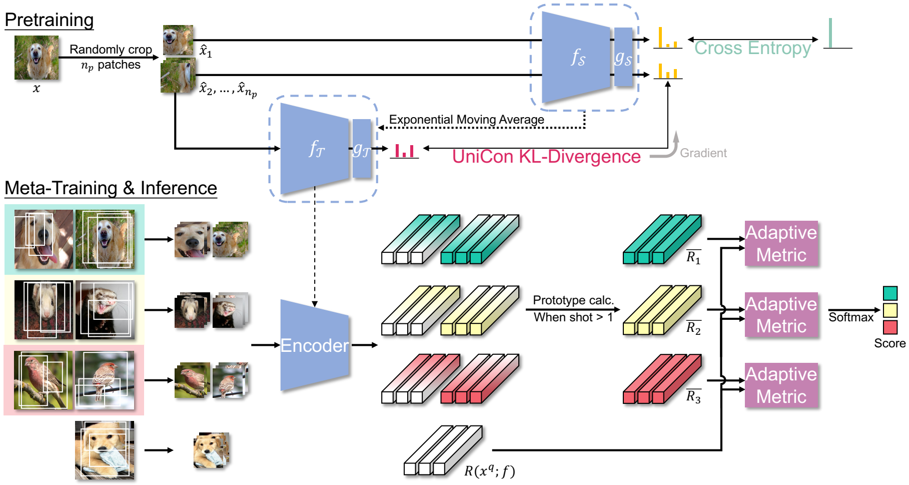

Fig. 2: Overview of our framework (3-way 2-shot as an example).

17, 22, 45]. Given a labeled dataset $D_{base} = \{(x_i^b, y_i^b)\}_{i=1}^{N_{base}}$ composed of n_c base classes, few-shot classification aims to construct a classifier for future tasks consisting of novel classes. 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:

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

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, $R(x_i^q; f)$ and $\overline{R_j}$[4] are defined as:

$$\begin{aligned} R(x_i^q; f) &:= \{\mathbf{u}_m | m = 1, 2, \ldots, n\}, \quad \overline{R_j} := \{\mathbf{v}_m | m = 1, 2, \ldots, n\}, \\ \text{where } \mathbf{u}_m &= f(\xi(x_i^q)), \quad \mathbf{v}_m = \frac{1}{K} \sum\nolimits_{i=1}^{NK} f(\xi(x_i^s)) \cdot [y_i^s = j], \end{aligned} \tag{2}$$

where $[y_i^s = j]$ is an indicator function that equals 1 when $y_i^s = j$ and 0 otherwise.

[4] For cases where shot $K > 1$, we conduct additional prototype calculation for a structured FC layer [45].

4 Feature Calibration

4.1 Feature Calibration with Soft Labels

Different from existing methods using only ground-truth hard labels, we propose a novel paradigm that takes soft labels into account as well for pretraining few-shot encoders. It calibrates the extracted features by avoiding false supervision while fully exploiting the class-level diversity of patches. 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 resize and flip). We reserve the first element $\hat{x}_1$ for normal hard label supervision using the cross-entropy loss:

$$\mathcal{L}_{CE} = -\mathbf{y}^\top \log\left(\sigma(\phi_{\mathcal{S}}(\hat{x}_1))\right), \tag{3}$$

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. Specifically, 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 [35]:

$$\theta_{\mathcal{T}}^i \leftarrow m\theta_{\mathcal{T}}^{i-1} + (1-m)\theta_{\mathcal{S}}^i, \tag{4}$$

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 [12, 13]. To align the output of $\phi_{\mathcal{S}}$ to that of $\phi_{\mathcal{T}}$, we propose a UniCon KL-Divergence as described below.

4.2 UniCon 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:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{n_c} \exp(z_j)}. \tag{5}$$

As a common choice to measure two probability distributions, KL-Divergence is often used for soft label supervision [11, 16, 46]. 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 (a). The probabilities of the i-th binary classification $\mathbf{b}_i = [q_i, q_{\neg i}]$ can be obtained by:

$$q_i = \frac{\exp(z_i)}{\sum_{j=i}^{n_c} \exp(z_j)}, \quad q_{\neg i} = \frac{\sum_{k=i+1}^{n_c} \exp(z_k)}{\sum_{j=i}^{n_c} \exp(z_j)}. \tag{6}$$

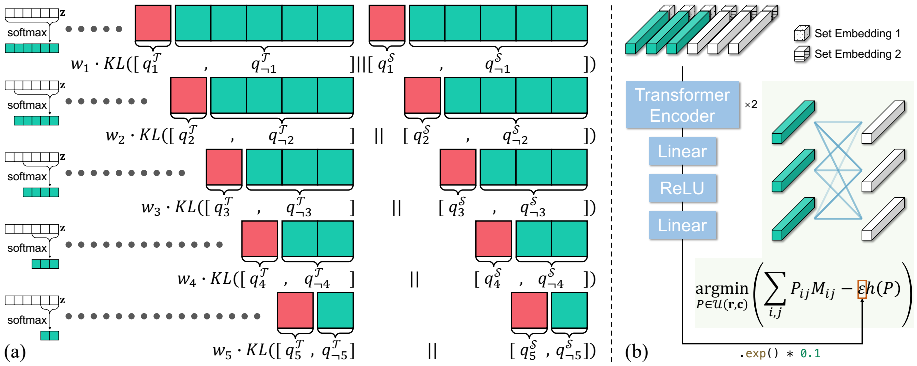
z
softmax
$w_1 \cdot KL([q_1^{\mathcal{T}}, q_{\neg 1}^{\mathcal{T}}] || [q_1^{\mathcal{S}}, q_{\neg 1}^{\mathcal{S}}])$
$w_2 \cdot KL([q_2^{\mathcal{T}}, q_{\neg 2}^{\mathcal{T}}] || [q_2^{\mathcal{S}}, q_{\neg 2}^{\mathcal{S}}])$
$w_3 \cdot KL([q_3^{\mathcal{T}}, q_{\neg 3}^{\mathcal{T}}] || [q_3^{\mathcal{S}}, q_{\neg 3}^{\mathcal{S}}])$
$w_4 \cdot KL([q_4^{\mathcal{T}}, q_{\neg 4}^{\mathcal{T}}] || [q_4^{\mathcal{S}}, q_{\neg 4}^{\mathcal{S}}])$
$w_5 \cdot KL([q_5^{\mathcal{T}}, q_{\neg 5}^{\mathcal{T}}] || [q_5^{\mathcal{S}}, q_{\neg 5}^{\mathcal{S}}])$
(a)
Set Embedding 1
Set Embedding 2
Transformer Encoder
×2
Linear
ReLU
Linear
$\underset{P \in \mathcal{U}(\mathbf{r},\mathbf{c})}{\text{argmin}} \left(\sum_{i,j} P_{ij} M_{ij} - \varepsilon h(P) \right)$
.exp() * 0.1
(b)

Fig. 3: Illustration of (a) the continuous binary classification process corresponding to the reformulation of KL-Divergence, and (b) the proposed Adaptive Metric formulating the measurement process as an OT problem. To handle various set compositions, the adjustment coefficient of an entropy regularization is tuned by a Modulate Module.

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 (the detailed proof can be found in the supplementary material).

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:*

$$KL(\mathbf{p}^{\mathcal{T}} || \mathbf{p}^{\mathcal{S}}) = \sum_{i=1}^{n_c - 1} w_i \cdot KL(\mathbf{b}_i^{\mathcal{T}} || \mathbf{b}_i^{\mathcal{S}}), \quad where\ w_i = \sum_{k=i}^{n_c} p_k^{\mathcal{T}}. \tag{7}$$

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

$$\tilde{w}_i = \frac{1}{n_c - i + 1} \sum_{k=i}^{n_c} p_k^{\mathcal{T}}. \tag{8}$$

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 the prior of normal intra-class classification tasks, *i.e.*, the input must belong to a certain base class. In this case, 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 [43].

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 [13] and knowledge distillation [16]. 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:

$$\begin{aligned}&\lim_{T\to\infty} |\tilde{w}_\alpha(T) - \tilde{w}_\beta(T)| \\ &= \lim_{T\to\infty} \left| \frac{\sum_{k_1=\alpha}^{n_c} \exp\left(z_{k_1}^{\mathcal{T}}/T\right)}{(n_c-\alpha+1)\sum_{j=1}^{n_c}\exp\left(z_j^{\mathcal{T}}/T\right)} - \frac{\sum_{k_2=\beta}^{n_c} \exp\left(z_{k_2}^{\mathcal{T}}/T\right)}{(n_c-\beta+1)\sum_{j=1}^{n_c}\exp\left(z_j^{\mathcal{T}}/T\right)} \right| = 0,\end{aligned} \tag{9}$$

according to which we derive a weighting scheme emphasizing the uniform contribution of different base classes for learning few-shot encoders:

$$w_i' := \lim_{T\to\infty} \frac{\sum_{k=i}^{n_c} \exp\left(z_k^{\mathcal{T}}/T\right)}{\sum_{j=1}^{n_c}\exp\left(z_j^{\mathcal{T}}/T\right)} = \frac{n_c - i + 1}{n_c}. \tag{10}$$

With a more rational weighting scheme, we define UniCon KL-Divergence that is used to compute the loss for soft labels $\mathcal{L}_{UKD}$:

$$UKD(\mathbf{p}^{\mathcal{T}}||\mathbf{p}^{\mathcal{S}}) := \sum_{i=1}^{n_c-1} w_i' \cdot KL(\mathbf{b}_i^{\mathcal{T}}||\mathbf{b}_i^{\mathcal{S}}), \tag{11}$$

$$\mathcal{L}_{UKD} = \frac{1}{n_p - 1}\sum_{i=2}^{n_p} UKD(\sigma(\phi_T(\hat{x}_i))||\sigma(\phi_S(\hat{x}_i))). \tag{12}$$

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

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{UKD}. \tag{13}$$

5 Adaptive Metric

After pretraining, $f_{\mathcal{T}}$ will be used as the feature extractor for further meta-training as illustrated in Fig. 2, where we propose Adaptive Metric for classification.

5.1 Review on Optimal Transport Distances

OT distances measure two sets by considering a hypothetical process of transporting goods from nodes of one set (suppliers) to nodes of the other set (demanders). Given the weight vectors $\mathbf{r}, \mathbf{c} \in \mathbb{R}_+^d$ ($\mathbf{r}^\top \mathbf{1}_d = \mathbf{c}^\top \mathbf{1}_d = 1$) where each element represents the total supply (demand) goods of a node, and the cost per unit M_{ij} for transporting from supplier i to demander j, the goal is to find a transportation plan with the lowest total cost from a set of valid plans $\mathcal{U}(\mathbf{r},\mathbf{c}) = \{P \in \mathbb{R}_+^{n\times n} | P\mathbf{1}_n = \mathbf{r}, P^\top \mathbf{1}_n = \mathbf{c}\}$. Based on the above notations, EMD [28] solves $\arg\min_{P\in\mathcal{U}(\mathbf{r},\mathbf{c})} \sum_{i,j} P_{ij}M_{ij}$ directly, which tends to produce a sparse P as it's usually solved on a vertex of the transport polytope.

5.2 The Sinkhorn Distance for Few-Shot Classification

To endow the algorithm with the ability to handle sets consisting of similar local features, we formulate the set measurement problem as optimizing an entropy-regularized OT problem proposed by [9] to encourage smoother transport matrices:

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

where $h(P) = -\sum_{i,j} P_{ij}\log P_{ij}$ is the information entropy of P and $\varepsilon \in (0,\infty)$ serves as an adjustment coefficient. $h(P)$ reflects the smoothness of P, the higher the entropy, the smoother the solved matrix.

Given two local feature sets to be measured, *i.e.*, $R(x_i^q; f)$ and $\overline{R_j}$, we define the weight of each feature with its cosine similarity to the mean of the other set, along with a softmax function to convert it to a probability distribution:

$$r_i = \frac{\exp(\hat{r}_i)}{\sum_{j=1}^n \exp(\hat{r}_j)}, \quad \text{where } \hat{r}_i := \frac{\mathbf{u}_i^\top \cdot \frac{1}{n}\sum_{j=1}^n \mathbf{v}_j}{\|\mathbf{u}_i\| \cdot \|\frac{1}{n}\sum_{j=1}^n \mathbf{v}_j\|}, \tag{15}$$

$$c_i = \frac{\exp(\hat{c}_i)}{\sum_{j=1}^n \exp(\hat{c}_j)}, \quad \text{where } \hat{c}_i := \frac{\mathbf{v}_i^\top \cdot \frac{1}{n}\sum_{j=1}^n \mathbf{u}_j}{\|\mathbf{v}_i\| \cdot \|\frac{1}{n}\sum_{j=1}^n \mathbf{u}_j\|}. \tag{16}$$

This stems from the intuition that local features more similar to the other set are more likely to be related to the concurrent foreground objects and, hence should be assigned greater weight. Then, with the cost to transport a unit from node $\mathbf{u}_i$ to $\mathbf{v}_j$ defined by their cosine similarity:

$$M_{ij} := 1 - \frac{\mathbf{u}_i^\top \mathbf{v}_j}{\|\mathbf{u}_i\|\|\mathbf{v}_j\|}, \tag{17}$$

we solve the optimization problem of Eq. (14) in parallel by the Sinkhorn-Knopp algorithm [32]. Eventually, with the solved P^ε, we define Adaptive Metric to compute the classification score that is used for cross entropy calculation or inference:

$$S(R(x_i^q; f_\mathcal{T}), \overline{R_j}) := \sum_{i=1}^n \sum_{j=1}^n (1 - M_{ij})P_{ij}^\varepsilon. \tag{18}$$

5.3 Modulate Module

Furthermore, to control the smoothness of the transport matrix adaptively according to specific local feature sets, we design a Modulate Module to predict

ε in Eq. (14) instead of treating it as a pre-fixed hyperparameter. 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. 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 [36], considering that its inductive bias suits the task of modeling the relationship between local features very well. As shown in Fig. 3 (b), 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 from the default value of 0.1.

The overall training process of our method is described in Algorithm 1.

Algorithm 1 Training process of FCAM.

```
PRETRAINING
Warm up $\phi_{\mathcal{S}}$ with $\mathcal{L}_{CE}$;
$\theta_{\mathcal{T}} \leftarrow \theta_{\mathcal{S}}$;
while epochs do
  while steps do
    Randomly crop $n_p$ patches $\{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{n_p}\}$ for each image $x$ in the minibatch;
    Calculate $\mathcal{L}_{CE}$ with $\hat{x}_1$;
    Calculate $\mathcal{L}_{UKD}$ with $\{\hat{x}_2, \hat{x}_3, \dots, \hat{x}_{n_p}\}$;
    $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{UKD}$;
    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
```

6 Experiments

Datasets. The experiments are conducted on three popular benchmarks: (1) ***mini*ImageNet** [37] is a subset of ImageNet [29] 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** [27] is also a subset of ImageNet [29] 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** [38] contains 200 bird categories with 11,788 images, which represents a fine-grained scenario. Following the splits in [41], 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 Modulate Module.

Training details. In the pretraining stage, we set $n_p = 4$ and $m = 0.999$. $\mathcal{L}_{UKD}$ 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 Modulate 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.

6.1 Comparison with State-of-the-art Methods

For general few-shot classification, we compare our method with the state-of-the-art methods in Tab. 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 performance. For fine-grained few-shot classification, we compare our method with the state-of-the-art methods in Tab. 2. Benefit from higher quality local features, the discriminative regions can be depicted more accurately, resulting in significant improvement against other methods, *i.e.*, **4.80**% and **3.03**% for 1-shot and 5-shot respectively against previous state-of-the-art method [22]. In particular, our method even outperforms state-of-the-art transductive [5,6] and cross-modal [6,18] methods, shedding some light on how much the poor local representations can degrade the performance in the fine-grained scenario.

6.2 Ablation Study

To begin with, a coarse-scale ablation is presented in Tab. 3. 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.

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. 4. We first pretrain 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 calibration. 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. 5 also illustrate that Feature Calibration results in better clusters for novel classes.

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† [37]	*ResNet12*	63.08 ± 0.80	75.99 ± 0.60	68.50 ± 0.92	80.60 ± 0.71
ProtoNet† [33]	*ResNet12*	60.37 ± 0.83	78.02 ± 0.57	65.65 ± 0.92	83.40 ± 0.65
TADAM [25]	*ResNet12*	58.50 ± 0.30	76.70 ± 0.30	-	-
FEAT [41]	*ResNet12*	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
DeepEMD [44]	*ResNet12*	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
Meta-Baseline [8]	*ResNet12*	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.74 ± 0.18
FRN [39]	*ResNet12*	66.45 ± 0.19	82.83 ± 0.13	72.06 ± 0.22	86.89 ± 0.14
PAL [23]	*ResNet12*	69.37 ± 0.64	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47
MCL [22]	*ResNet12*	69.31 ± 0.20	85.11 ± 0.20	73.62 ± 0.20	86.29 ± 0.20
DeepEMD v2 [45]	*ResNet12*	68.77 ± 0.29	84.13 ± 0.53	74.29 ± 0.32	87.08 ± 0.60
FADS [31]	*ResNet12*	66.73 ± 0.88	83.51 ± 0.51	74.12 ± 0.74	86.56 ± 0.46
Centroid Alignment‡ [1]	*WRN-28-10*	65.92 ± 0.60	82.85 ± 0.55	74.40 ± 0.68	86.61 ± 0.59
Oblique Manifold‡ [26]	*ResNet18*	63.98 ± 0.29	82.47 ± 0.44	70.50 ± 0.31	86.71 ± 0.49
FewTURE‡ [15]	*ViT-Small*	68.02 ± 0.88	84.51 ± 0.53	72.96 ± 0.92	86.43 ± 0.67
FCAM (ours)	*ResNet12*	**70.20 ± 0.28**	**85.61 ± 0.52**	**75.02 ± 0.31**	**88.41 ± 0.59**

† results are reported in [45]. ‡ 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† [37]	*ResNet12*	71.87 ± 0.85	85.08 ± 0.57
ProtoNet† [33]	*ResNet12*	66.09 ± 0.92	82.50 ± 0.58
DeepEMD [44]	*ResNet12*	75.65 ± 0.83	88.69 ± 0.50
FRN♯ [39]	*ResNet12*	78.86 ± 0.28	90.48 ± 0.16
MCL♯ [22]	*ResNet12*	79.39 ± 0.29	90.48 ± 0.49
DeepEMD v2 [45]	*ResNet12*	79.27 ± 0.29	89.80 ± 0.51
Centroid Alignment‡ [1]	*ResNet18*	74.22 ± 1.09	88.65 ± 0.55
Oblique Manifold‡ [26]	*ResNet18*	78.24 ± −	92.15 ± −
ECKPN♭ [5]	*ResNet12*	77.43 ± 0.54	92.21 ± 0.41
AGAM♮ [18]	*ResNet12*	79.58 ± 0.25	87.17 ± 0.23
ADRGN♭♮ [6]	*ResNet12*	82.32 ± 0.51	92.97 ± 0.35
FCAM (ours)	*ResNet12*	**83.20 ± 0.27**	**93.22 ± 0.39**

† results are reported in [45].
‡ methods with bigger backbones.
♯ reproduced using the data split we use.
♭ transductive methods.
♮ methods that use attribute information.
The second best results are underlined.

Table 3: Ablation of Feature Calibration and Adaptive Metric. The experiments are conducted with *ResNet12* on *mini*ImageNet.

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
✓	✓	**70.20 ± 0.28**	**85.61 ± 0.52**

Table 4: Comparison of using classical and UniCon KL-Divergence for calibration (top), and the results of whether using Modulate Module to adjust ε (bottom).

Setting	1-shot	5-shot
Classical KL-Divergence	69.94 ± 0.28	84.79 ± 0.53
UniCon KL-Divergence	**70.20 ± 0.28**	**85.61 ± 0.52**
w/o Modulate Module	69.69 ± 0.28	85.23 ± 0.52
w/ Modulate Module	**70.20 ± 0.28**	**85.61 ± 0.52**

UniCon KL-Divergence is more suitable for Feature Calibration. We compare different temperature settings in Fig. 4. 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 UniCon KL-Divergence, *i.e.*, $T \to \infty$, constantly outperforms other settings. Furthermore, UniCon KL-Divergence yields better final performance than classical KL-Divergence as shown in Tab. 4. Both the above experiments demonstrate the importance of a smoother weighting scheme in Feature Calibration.

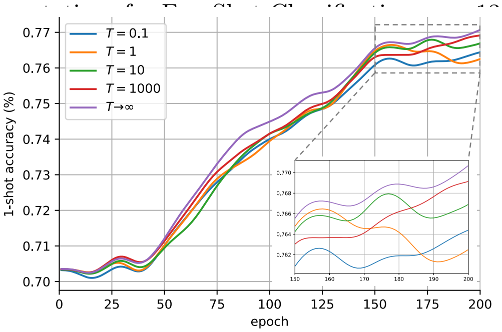

Fig. 4: Gaussian smoothed 1-shot test accuracy curves on CUB-200-2011 during feature calibration, with different temperatures to adjust the weighting scheme of the classical KL-Divergence. The results of the same 1000 tasks are averaged for each data point.

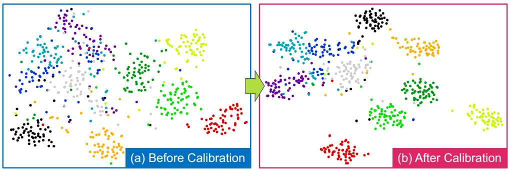

Fig. 5: Visualization [24] of novel class samples embedded by encoders trained (a) without and (b) with Feature Calibration.

Entropic term handles sets consisting of similar nodes. For sets consisting of similar local features, the transport matrix solved by EMD (Fig. 6 (a)) is very sparse, which tries to match a feature with few "most" similar opposite features. In contrast, Adaptive Metric (Fig. 6 (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 of them by allowing "one-to-many" matching.

Modulate Module brings adaptability. For sets consisting of similar local features, Modulate Module predicts a relatively larger ε, resulting in a smoother transport matrix (Fig. 6 (b)). For sets consisting of dissimilar local features, a relatively smaller ε is produced, making the transport matrix moderately sparse (Fig. 6 (c)). Quantitative results of whether using Modulate Module to adjust ε is also presented in Tab. 4. Compared to a fixed default value, it introduces flexibility into the metric process, helping achieve higher performance by realizing an adaptive measurement.

6.3 Cross-Domain Experiments

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. This 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 Tab. 5, our method outperforms the previous state-of-the-art approach, demonstrating its superiority in improving novel-class generalization.

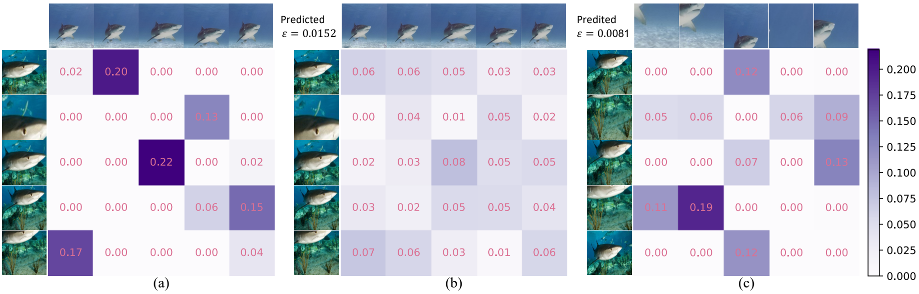

Fig. 6: Visualization of solved transport matrices. Results of (a) EMD and (b) Adaptive Metric for sets consisting of similar local features, and the result of (c) Adaptive Metric for sets consisting of dissimilar local features.

Table 5: Cross-domain experiments (*mini*ImageNet→CUB) following the setting of [7]. Average 5-way 1-shot and 5-way 5-shot accuracy (%) with 95% confidence intervals.

Method	1-shot	5-shot
ProtoNet† [33]	50.01 ± 0.82	72.02 ± 0.67
MatchNet† [37]	51.65 ± 0.84	69.14 ± 0.72
cosine classifier [7]	44.17 ± 0.78	69.01 ± 0.74
linear classifier [7]	50.37 ± 0.79	73.30 ± 0.69
KNN [21]	50.84 ± 0.81	71.25 ± 0.69
DeepEMD v2 [45]	$\underline{54.24 \pm 0.86}$	$\underline{78.86 \pm 0.65}$
FCAM (ours)	$\mathbf{58.20 \pm 0.30}$	$\mathbf{80.92 \pm 0.65}$

† results are reported in [45].
The second best results are underlined.

Table 6: Results of global-representation-based FSC methods on *mini*ImageNet, w/o and w/ Feature Calibration (FC). Average 5-way 1-shot and 5-way 5-shot accuracy (%) with 95% confidence intervals.

Method	Setting	1-shot	5-shot
cosine classifier [7]	w/o FC	61.31 ± 0.20	77.73 ± 0.21
	w/ FC	$\mathbf{64.92 \pm 0.20}$	$\mathbf{80.51 \pm 0.21}$
linear classifier [7]	w/o FC	55.74 ± 0.20	78.89 ± 0.21
	w/ FC	$\mathbf{59.35 \pm 0.20}$	$\mathbf{81.46 \pm 0.20}$
Classifier-Baseline [8]	w/o FC	60.67 ± 0.21	78.53 ± 0.21
	w/ FC	$\mathbf{64.33 \pm 0.21}$	$\mathbf{81.01 \pm 0.21}$
Meta-Baseline [8]	w/o FC	63.62 ± 0.21	80.25 ± 0.20
	w/ FC	$\mathbf{64.90 \pm 0.21}$	$\mathbf{81.04 \pm 0.21}$

6.4 Feature Calibration for Global-Representation-based FSC

Although Feature Calibration is proposed for improving local representations, it also benefits methods based on global representations as demonstrated in Tab. 6. Feature Calibration boosts the performance of these methods significantly due to its ability to leverage the class-level diversity provided by random cropping.

7 Conclusion

In this paper, we presented a novel FCAM method for few-shot classification to unleash the power of local representations in improving novel-class generalization. It improves the few-shot encoder by calibrating it towards the test scenario and handles various set compositions of local feature sets adaptively. Our method achieves new state-of-the-art performance on multiple datasets.

References

1. Afrasiyabi, A., Lalonde, J.F., Gagn'e, C.: Associative alignment for few-shot image classification. In: European Conference on Computer Vision (ECCV). pp. 18–35. Springer (2020)
2. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Advances in Neural Information Processing Systems. vol. 27 (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf
3. Buciluundefined, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 535–541. KDD '06 (2006). https://doi.org/10.1145/1150402.1150464, https://doi.org/10.1145/1150402.1150464
4. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3) (Jun 2011). https://doi.org/10.1145/1970392.1970395, https://doi.org/10.1145/1970392.1970395
5. Chen, C., Yang, X., Xu, C., Huang, X., Ma, Z.: 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). pp. 6596–6605 (June 2021)
6. Chen, C., Yang, X., Yan, M., Xu, C.: Attribute-guided dynamic routing graph network for transductive few-shot learning. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 6259–6268. MM '22 (2022). https://doi.org/10.1145/3503161.3548301, https://doi.org/10.1145/3503161.3548301
7. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
8. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: Exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9062–9071 (October 2021)
9. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems. vol. 26 (2013), https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf
10. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135 (06–11 Aug 2017), https://proceedings.mlr.press/v70/finn17a.html
11. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1607–1616 (10–15 Jul 2018), https://proceedings.mlr.press/v80/furlanello18a.html
12. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 21271–21284 (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

14. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
15. Hiller, M., Ma, R., Harandi, M., Drummond, T.: Rethinking generalization in few-shot classification. In: Advances in Neural Information Processing Systems (NeurIPS) (2022), https://openreview.net/forum?id=p_g2nHlMus
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv e-prints arXiv:1503.02531 (Mar 2015). https://doi.org/10.48550/arXiv.1503.02531
17. Hu, S.X., Li, D., Stühmer, J., Kim, M., Hospedales, T.M.: 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). pp. 9068–9077 (June 2022)
18. Huang, S., Zhang, M., Kang, Y., Wang, D.: 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). https://doi.org/10.1609/aaai.v35i9.16957, https://ojs.aaai.org/index.php/AAAI/article/view/16957
19. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. In: Advances in Neural Information Processing Systems. vol. 31 (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/6d9cb7de5e8ac30bd5e8734bc96a35c1-Paper.pdf
20. Li, H., Dong, W., Mei, X., Ma, C., Huang, F., Hu, B.G.: LGM-net: Learning to generate matching networks for few-shot learning. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3825–3834 (09–15 Jun 2019), https://proceedings.mlr.press/v97/li19c.html
21. Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
22. Liu, Y., Zhang, W., Xiang, C., Zheng, T., Cai, D., He, X.: Learning to affiliate: Mutual centralized learning for few-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14411–14420 (June 2022)
23. Ma, J., Xie, H., Han, G., Chang, S.F., Galstyan, A., Abd-Almageed, W.: Partner-assisted learning for few-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10573–10582 (October 2021)
24. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
25. Oreshkin, B., Rodríguez López, P., Lacoste, A.: Tadam: Task dependent adaptive metric for improved few-shot learning. In: Advances in Neural Information Processing Systems. vol. 31 (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/66808e327dc79d135ba18e051673d906-Paper.pdf
26. Qi, G., Yu, H., Lu, Z., Li, S.: Transductive few-shot classification on the oblique manifold. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8412–8422 (October 2021)
27. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classifica-

tion. In: Proceedings of 6th International Conference on Learning Representations ICLR (2018)

28. Rubner, Y., Guibas, L., Tomasi, C.: The earth mover"s distance, multidimensional scaling, and color-based image retrieval. Proceedings of the Arpa Image Understanding Workshop (1997)
29. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
30. Schwartz, E., Karlinsky, L., Shtok, J., Harary, S., Marder, M., Kumar, A., Feris, R., Giryes, R., Bronstein, A.: Delta-encoder: an effective sample synthesis method for few-shot object recognition. In: Advances in Neural Information Processing Systems. vol. 31 (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/1714726c817af50457d810aae9d27a2e-Paper.pdf
31. Shao, S., Wang, Y., Liu, B., Liu, W., Wang, Y., Liu, B.: Fads: Fourier-augmentation based data-shunting for few-shot classification. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2023). https://doi.org/10.1109/TCSVT.2023.3292519
32. Sinkhorn, R., Knopp, P.: Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics **21**(2), 343–348 (1967)
33. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. vol. 30 (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf
34. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
35. Tarvainen, A., Valpola, H.: 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), https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
37. Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. vol. 29 (2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf
38. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
39. Wertheimer, D., Tang, L., Hariharan, B.: Few-shot classification with feature map reconstruction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8012–8021 (June 2021)
40. Xiao, K.Y., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=gl3D-xY7wLq
41. Ye, H.J., Hu, H., Zhan, D.C., Sha, F.: 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)

42. Yim, J., Joo, D., Bae, J., Kim, J.: 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)
43. Yue, Z., Zhang, H., Sun, Q., Hua, X.S.: Interventional few-shot learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 2734–2746 (2020), https://proceedings.neurips.cc/paper/2020/file/1cc8a8ea51cd0adddf5dab504a285915-Paper.pdf
44. Zhang, C., Cai, Y., Lin, G., Shen, C.: 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)
45. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Differentiable earth mover's distance for few-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–17 (2022). https://doi.org/10.1109/TPAMI.2022.3217373
46. Zhang, Z., Sabuncu, M.: Self-distillation as instance-specific label smoothing. In: Advances in Neural Information Processing Systems. vol. 33, pp. 2184–2195 (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1731592aca5fb4d789c4119c65c10b4b-Paper.pdf