# Bilaterally-normalized Scale-consistent Sinkhorn Distance for Few-shot Image Classification

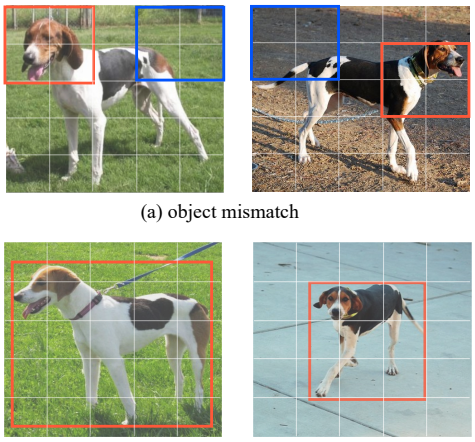
Yanbin Liu, Linchao Zhu, *Member, IEEE*, Xiaohan Wang
Makoto Yamada, and Yi Yang, *Senior Member, IEEE*

***Abstract*—Few-shot image classification aims at exploring transferable features from base classes to recognize images of the unseen novel classes with only a few labelled images. Existing methods usually compare the support features and query features, which are implemented by either matching the global feature vectors or matching the local feature maps at the same position. However, the few labelled images fail to capture all the diverse context and intra-class variations, leading to mismatch issues for existing methods. On one hand, due to the misaligned position and cluttered background, existing methods suffer from the *object mismatch* issue (Fig. 1(a)). On the other hand, due to the scale inconsistency between images, existing methods suffer from the *scale mismatch* issue (Fig. 1(b)). In this paper, we propose the Bilaterally-normalized Scale-consistent Sinkhorn Distance (BSSD) to solve these issues. Firstly, instead of same-position matching, we utilize the Sinkhorn Distance to find an optimal matching between images, mitigating the object mismatch caused by misaligned position. Meanwhile, we propose the intra-image and inter-image attentions as the bilateral normalization on Sinkhorn Distance to suppress the object mismatch caused by background clutter. Secondly, local feature maps are enhanced with the multi-scale pooling strategy, making Sinkhorn Distance possible to find a consistent matching scale between images. Experimental results show the effectiveness of the proposed approach, and we achieve the state-of-the-art on three few-shot benchmarks.**

***Index Terms*—Few-shot, Sinkhorn Distance, Attention**

## I. Introduction

OVER the past few years, convolutional neural networks (CNNs) have achieved tremendous breakthroughs in a wide range of computer vision tasks, such as image classification [1], object detection [2], or semantic segmentation [3]. Usually, it requires large-scale datasets to effectively train a network. Collecting such large amounts of data is extremely laborious and time-consuming, while in some cases like rare species recognition or fine-grained classification [4], it is even infeasible. On the contrary, the human visual system has the ability to learn novel concepts with only one or few instances [5]. For example, humans (even children) can acquire the concept of "banana" after seeing a single banana image and be able to recognize future bananas. Mimicking this generalized learning ability of humans, few-shot learning (FSL) aims to recognize instances of the unseen novel concepts (*i.e.*, *query set*) with only few labeled instances (*i.e.*, *support set*) by exploring the latent patterns from the available seen concepts.



(a) object mismatch

(b) scale mismatch

Fig. 1. Few-shot feature matching encounters two problems: (a) object mismatch, *i.e.*, the misaligned position (orange box) and background clutter (blue box) will prevent the correct matching; (b) scale mismatch, *i.e.*, the scale inconsistency between images will cause incorrect matching.

To tackle the few-shot learning problem, a variety of algorithms have been proposed. These algorithms can be divided into three types, *i.e.*, metric-based [6]–[14], gradient-based [15]–[19], and transfer-based [20]–[23] methods. *Metric-based methods* aim to learn a generalizable feature space and utilize a distance function (such as Cosine or Euclidean) to compute the similarities between support and query images. *Gradient-based methods* try to learn a task-level meta-learner that can quickly adapt model parameters to a new task with few examples. *Transfer-based methods* consider a simple transfer learning baseline by first pre-training a model on the large meta-training classes. Then, a classifier is trained on the meta-test classes to utilize the transferable knowledge from the pre-trained model.

Early few-shot methods follow the common practice of large-scale image classification [1] to extract global feature vectors from CNNs. However, global feature vectors often rely on a large amount of training data to cover diverse context

This work was supported in part by the Australian Research Council (ARC) under Grant DP200100938. M.Y. was supported by MEXT KAKENHI 20H04243 and partly supported by MEXT KAKENHI 21H04874. (*Corresponding author: Yi Yang.*)

Yanbin Liu is with the School of Computing, Australian National University, Canberra, ACT 2601, Australia. (email: csyanbin@gmail.com)

Linchao Zhu, Xiaohan Wang, and Yi Yang are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. (e-mail: zhulinchao@zju.edu.cn, xiaohan.wang@zju.edu.cn, yangyics@zju.edu.cn)

Makoto Yamada is with Okinawa Institute of Science and Technology, Okinawa 904-0495, Japan. (email: makoto.yamada@oist.jp)

and intra-class variations, which is infeasible in the few-shot problem. To ameliorate global feature vectors, a natural way is to employ the local feature maps (*e.g.*, by removing the last pooling layer in CNNs) [24]–[29]. Compared with global feature vectors, local feature maps extract more meticulous details of the image, thus amplifying the global feature vectors with more discriminative and transferable information. Existing few-shot methods [26], [27] adopt local feature maps in a position-to-position manner, *i.e.*, matching the corresponding feature maps at the same position and aggregating the results. Although this is simple and straightforward, due to the limited data constraints in few-shot learning, it suffers from two practical issues: object mismatch and scale mismatch (Fig. 1).

First, *object mismatch* refers to the issue that object parts of the same category are not correctly matched. For example, in Fig. 1(a), for a position-to-position matching method, the left dog's head (orange box in the left image) will match the background (blue box in the right image) rather than the correct right dog's head (orange box in the right image). The object mismatch issue is caused by the misaligned object position and cluttered background. Current local feature based methods do not address this issue well. For example, DC [27] learns a single classifier for each position without alignment, thus ignoring the mismatch issue. DN4 [26] only considers the top-$k$ matching in a class, which is sub-optimal for dealing with the mismatch issue and it can be interfered by the irrelevant background features. In this paper, we propose a Bilaterally-normalized Sinkhorn Distance to solve the object mismatch. At first, we model the local feature matching problem in a Sinkhorn distance framework instead of the position-to-position matching. Specifically, we calculate the pairwise matching costs between the local feature maps of the support image and query image. Feature matching is formulated as a total costs minimization problem, which can be solved efficiently by the Sinkhorn iteration. This way, the optimal matching takes into account the whole feature maps to correct the misaligned positions. Moreover, we devise the intra-image and inter-image attentions for each position to serve as the bilateral normalization on Sinkhorn distance. This normalization explicitly diminishes the effect of the background features and enhances the weights of the object features, mitigating the object-to-background mismatch.

Second, *scale mismatch* refers to the issue that the objects from the same category have inconsistent scales across various images. For example, in Fig. 1(b), the dog in the left image is approximately twice the scale of the right one. Scale mismatch would cause adverse effects (*e.g.*, degraded performance for small objects) in various tasks, such as object detection [30], image segmentation [31]. However, it is seldom studied in few-shot image classification. To solve this issue, we first apply the multi-scale pooling to the original feature map of the support image to generate local features representing multiple scales. These local features are concatenated to replace the original ones in the Sinkhorn distance. Then, we keep the query features unchanged and compute the Bilaterally-normalized Sinkhorn distance to find the optimal matching between the query features and the multi-scale support features. This strategy makes Sinkhorn distance possible to find and utilize the consistent matching scale between support and query images. Even when different object parts are disproportionately scaled (non-rigid deformation), the concatenation makes it feasible to find the correct parts combination across various scales.

All the proposed modules are incorporated in a general framework to compute the distance between the support and query images (Fig. 2). We refer to this distance as **Bilaterally-normalized Scale-consistent Sinkhorn Distance (BSSD)**. In the experiments, we show the effectiveness of each component of the proposed model. Our main contributions are summarized as follows:

- We propose a novel Bilaterally-normalized Scale-consistent Sinkhorn Distance (BSSD) to obtain a reliable local feature matching for few-shot classification.
- We devise the intra-image and inter-image attentions as the bilateral normalization on Sinkhorn distance to diminish the cluttered background issue.
- The proposed Bilaterally-normalized Scale-consistent Sinkhorn Distance is capable of finding the consistent matching scale between images, thus addressing the scale mismatch issue.
- We demonstrate the effectiveness and generalizability of the proposed method via extensive experiments and achieve the state-of-the-art on three benchmark datasets.

## II. Related Work

### A. Few-shot Image Classification

In few-shot image classification, we are given abundant examples from the base classes (seen), while the goal is to learn to recognize novel classes (unseen) with few labeled examples. Much efforts have been devoted to deal with this task from three different views, *i.e.*, metric-based methods, gradient-based methods, and transfer-based methods.

**Metric-based methods** try to learn common image features as well as a non-parametric classifier (*e.g.*, Euclidean or Cosine distance) that can generalize from base classes to novel classes. Matching Network [6] introduces the concept of support/query set and $N$-way $K$-shot learning protocol. The Cosine similarity is calculated between one query image and all support images. Prototypical Network [7] computes the Euclidean distance between the query feature and the prototypes of all class (a prototype refers to the mean of support features belonging to the same class). Except for the simple distances, neural networks can serve as a learnable distance function. Relation Network [32] learns a deep distance metric from the concatenation of query and support representations. FEAT [14] utilizes a set-to-set transformation to make both the features and the distances task-specific. The proposed Bilaterally-normalized Scale-consistent Sinkhorn distance shares the same spirit of learning a discriminative distance metric. However, instead of using global feature vectors, we adopt the local feature maps that is more consistent between seen and unseen classes to mitigate the data-scarcity issue of few-shot learning.

**Gradient-based methods** aim to learn a general updating rule that can quickly adapt model parameters to tackle a new task with only few examples. MAML [15] proposes a model-agnostic gradient updating rule to find a good intialization

that can adapt quickly to any novel task with a few steps. To avoid the computation of the costy Hessian matrix in MAML, Reptile [16] employees only first-order derivatives for parameter updates. MetaOptNet [8] backpropagates though the optimization of the SVM classifier while avoids second-order derivatives in the feature backbone.

**Transfer-based methods** learn a pre-trained or self-supervised representation on the entire base classes and then train a classifier on top of this representation. Tian *et al.* [20] directly learn a linear classifier on top of a pre-trained embedding and use self-distillation for further improvement. Meta-Baseline [33] verifies the effect of both classifier pre-training and meta-learning.

**Generalized few-shot learning.** Besides the above few-shot learning methods, recent works start to discuss more generalized problems, such as domain-agnostic recognition [34], [35], utilizing extra semantic information [36], few-shot human–object interaction (HOI) [37] and novel discovery [38], [39]. For example, DGIG-Net [37] learns a task-oriented cross-modal graph with a novel graph prototypes framework for few-shot HOI. Our work is orthogonal to these generalized settings and can be adapted for their problems.

### B. Local Image Features

Local feature maps contain rich and meticulous information, making it suitable for various computer vision tasks, such as object detection [2], visual retrieval [40], semantic segmentation [3], or fine-grained classification [4]. These tasks make use of the spatial relations and detailed part representations for accurate image-level or pixel-level prediction. Inspired by their success, few-shot learning can benefit from the local feature maps to enrich the global feature vectors which are not well exploited under limited training examples. Actually, the local part representations are more consistent and transferable between the seen and unseen classes. DC [27] trains a classifier for each position in the feature maps but ignores the feature alignment. TFH [28] hallucinates the tensor features as data augmentation and performs global average pooling (GAP) for classification. DN4 [26] finds the top-$k$ matching descriptors between a query image and a class to calculate the image-to-class similarity. Compared with these methods, the proposed algorithm considers the whole local feature maps to find the global optimal matching, leading to better feature alignment.

For local feature maps, attention techniques [41] such as class activation maps (CAM) [42] are widely-used to highlight discriminative areas. In few-shot learning, SAML [43] modifies this activation by calculating the norm of local feature maps to suppress the semantically irrelevant local regions. [44] proposes a loss function CAM-loss to boost the classification performance in CNNs. We go a step further to compute both the intra-image and inter-image attentions, suppressing the clustered background of image pairs. Another good practice for local descriptors is incorporating multi-scale information, *e.g.*, FCN [3] and FPN [2]. In our method, the multi-scale pooling is applied to the support feature map to obtain features of multiple scales. These support features are concatenated to match the original query features. The proposed strategy is simple but non-trivial, since the concatenation offers a way of matching local features across all scales to deal with the disproportional object parts caused by non-rigid deformation.

### C. Optimal Transport

Optimal transport aims at computing a minimal cost transportation between a source distribution and a target distribution. Recently, it has been applied in various applications, such as image localization [45], domain adaptation [46], [47], generative model [48] and graph matching [49], [50]. The original optimal transport problem (a.k.a. Earth Mover's Distance) can be solved with network simplex or interior point methods, both of which require cubic complexity. In this paper, we formulate local feature matching as an entropy regularized version of the optimal transport problem (*i.e.*, Sinkhorn distance [51]), which only needs quadratic complexity. Moreover, all the proposed components (*i.e.*, optimal local feature matching, bilateral normalization, scale-consistency) can be incorporated in the Bilaterally-normalized Scale-consistent Sinkhorn distance (BSSD) to form a unified end-to-end framework.

## III. Methodology

This section presents the proposed framework. We first introduce some notations and formalize the few-shot classification problem. Then, we describe the Bilaterally-normalized Scale-consistent Sinkhorn Distance (BSSD) framework including optimal local feature matching, bilateral normalization, and scale-consistency. Finally, we utilize the Sinkhorn algorithm to optimize the local feature matching problem.

### A. Problem Definition

In this paper, we consider the standard few-shot classification setting. At first, we are given a labeled dataset consisting of base classes $C_{base}$ (a.k.a. seen classes) with an abundant number of images in each class. Then, the objective is to classify images in novel classes $C_{novel}$ (a.k.a. unseen classes) with only few images in each class. Since the base classes and novel classes have no overlap, *i.e.*, $C_{base} \cap C_{novel} = \emptyset$, it is difficult to directly fine-tune or transfer from the model learned on base classes $C_{base}$ [6], [15].

In order to break the non-overlapping constraint of base/novel classes, an episodic-training paradigm was proposed in [6]. Instead of optimizing over mini-batches of training examples $(\mathbf{x}_i, y_i)_{i=1}^{B}$, episodic-training learns over mini-batches of training tasks $\mathcal{T} = \{(D_i^{train}, D_i^{test})\}_{i=1}^{B}$. Here, $D_i^{train}$ and $D_i^{test}$ represent the training set and test set of the $i$-th task. Specifically, $D_i^{train}$ is called the *support set*, which contains $K$ images in each of the $N$ classes sampled from $C_{base}$. $D_i^{test}$ is called the *query set*, which contains $Q$ images in the same $N$ classes as the support set. $(D_i^{train}, D_i^{test})$ is denoted as an $N$-way $K$-shot few-shot task. With episodic-training, the model tries to generalize across multiple tasks rather than fitting a specific classifier on $C_{base}$, thus breaking the non-overlapping constraint. Due to this good property, episodic-training has been widely-used in recent few-shot learning methods [8], [14], [15], [26], [27], [52]. Therefore, we also follow this good practice in the proposed algorithm.

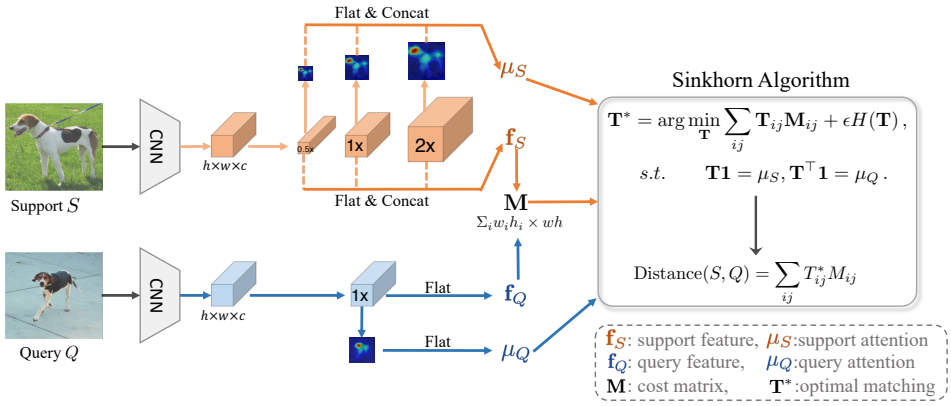



Fig. 2. Our framework for Sinkhorn distance computation. Given a support image $\mathbf{x}_S$ and a query image $\mathbf{x}_Q$, the local feature maps are first extracted with the CNN. For the support feature map, multi-scale pooling is utilized to generate multi-scale local features while the query feature map keeps unchanged. Then, the intra-image and inter-image attentions (section III-B3) are calculated as bilateral normalization that re-weights the object and background features. Finally, the concatenated features and attentions are incorporated in a unified optimization framework to calculate the Sinkhorn distance.

However, even equipped with episodic-training, few-shot models still suffer from the data scarcity issue since for each training task, only few (*e.g.*, $K = 1$ or 5) images can be used in each class. Few labelled images fail to represent the complex intra-class variations, resulting the object mismatch and scale mismatch problems. To tackle these problems, we model few-shot feature matching as an optimal local feature matching problem and then enforce the bilateral normalization as well as scale-consistency to address the mismatch problems.

### B. Bilaterally-normalized Scale-consistent Sinkhorn Distance

*1) Preliminary:* Our algorithm belongs to the metric-based few-shot learning methods, which learn a discriminative metric space by comparing the support and query images. While existing methods extract global feature vector for direct comparison, we propose to extract local feature maps and utilize the Sinkhorn distance for comparison, as shown in Fig. 2.

Given an image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, the convolution neural network $f_\theta$ extracts the feature map as $f_\theta(\mathbf{x}) \in \mathbb{R}^{h\times w\times c}$, where $\theta$ is the parameter of the network, $h, w$ are the height/width of the map, and $c$ is the feature dimension. Here, $f_\theta(\mathbf{x})$ can be reshaped to $f_\theta(\mathbf{x}) \in \mathbb{R}^{m\times c}$ ($m = hw$) and viewed as a set of $c$-dimensional ***local features***. Given the support image $\mathbf{x}_S$ and query image $\mathbf{x}_Q$, the key issue is how to compare $f_\theta(\mathbf{x}_S)$ and $f_\theta(\mathbf{x}_Q)$ with a proper distance.

A straightforward extension from global feature methods [6], [7] is to sum up the distances of local features from the same position as follows:

$$\mathrm{Distance}(\mathbf{x}_S, \mathbf{x}_Q) = \sum_{i=1}^{m} D(f_\theta(\mathbf{x}_S)[i], f_\theta(\mathbf{x}_Q)[i]), \quad (1)$$

where $D$ is a distance metric, such as Euclidean. Eqn. (1) is similar to DC [27], while the difference is that DC trains a classifier for each descriptor rather than direct distance computation. Although simple, Eqn. (1) and DC would suffer from the object mismatch and scale mismatch issues as shown in Fig. 1. DN4 [26] improves the distance of Eqn. (1) with $k$-NN. For each feature $q_i$ from $f_\theta(\mathbf{x}_Q)$, its $k$-nearest neighbors from $f_\theta(\mathbf{x}_S)$ are searched to compute the distance. Since DN4 only considers $k$ support descriptors for each query feature, it is suboptimal and may be sensitive to the choice of $k$.

*2) Optimal local feature matching:* Instead of computing the same position distance or $k$-nearest neighbor distance, we seek for a global optimal matching to calculate the local feature distance. At first, a cost matrix $\mathbf{C} \in \mathbb{R}^{m\times m}$ is calculated as $\mathbf{C}_{ij} = D(f_\theta(\mathbf{x}_S)[i], f_\theta(\mathbf{x}_Q)[j])$ with $\mathbf{C}_{ij}$ denoting the cost of matching the $i$-th support feature and the $j$-th query feature. The objective is to find an optimal matching $\mathbf{M}^*$ that can minimize the total cost of all pairwise matching:

$$\mathbf{M}^* = \arg\min_{\mathbf{M}} \sum_{ij} \mathbf{M}_{ij}\mathbf{C}_{ij},$$
$$s.t. \quad \mathbf{M} \geq \mathbf{0}, \mathbf{M1} = \mathbf{p}_S, \mathbf{M}^\top \mathbf{1} = \mathbf{p}_Q, \quad (2)$$

where $\mathbf{1} \in \mathbb{R}^{m\times 1}$ is a vector of all ones, $\mathbf{p}_S$ and $\mathbf{p}_Q$ are the support and query probabilities to restrict $\mathbf{M}$ to nontrivial solutions. The problem (2) is an optimal transport problem [53].

**Theorem 1** (Proposition 2.1 in [53]) *For problem (2), if* $\mathbf{p}_S = \mathbf{p}_Q = \frac{1}{m}\mathbf{1}$*, there exists a solution* $\mathbf{M}^*$*, which is a permutation matrix (the corresponding permutation* $\pi^* \in \mathrm{Perm(m)}$*).*

According to Theorem 1, problem (2) finds an **optimal alignment $\pi^*$ between the support and query features**. This way, the same parts (*e.g.*, head) of the same object in two images are well-aligned, thus solving the object mismatch issue caused by the misaligned object position. In contrast, existing methods either utilize same-position matching (DC [27]) or apply top-$k$ matching (DN4 [26]), failing to find the optimal matching.

*3) Bilateral normalization:* In Theorem 1, $\mathbf{p}_S$ and $\mathbf{p}_Q$ are assumed to be uniform distributions, which means that the object and background features are treated equally. Therefore, the cluttered background will interfere the correct object matching. To alleviate this, we first compute the ***intra-image attention*** with the extracted local features as follows:

$$\mathbf{p}_S^{\mathrm{intra}}[i] = \frac{\exp(\langle f_\theta(\mathbf{x}_S)[i], f_\theta(\mathbf{x}_S)[i]\rangle^{1/2})}{\sum_{j=1}^{m} \exp(\langle f_\theta(\mathbf{x}_S)[j], f_\theta(\mathbf{x}_S)[j]\rangle^{1/2})}. \quad (3)$$

Here, $\langle\ ,\ \rangle$ denotes inner product. In a well-trained CNN, the inner product of local features can serve as an indicator of

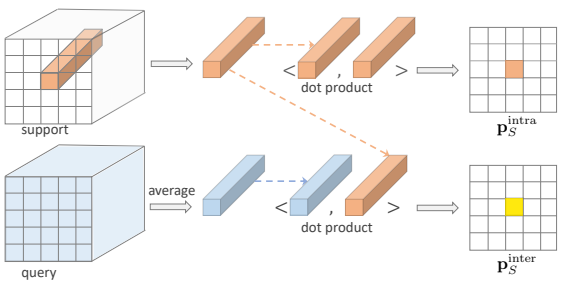



Fig. 3. Illustration of the intra-image and inter-image attentions.

object/background positions. This is because during training time the object positions are more related to the loss function, thus being trained to have larger feature values.

However, since the model is trained on base classes $C_{base}$ and applied to unseen novel classes $C_{novel}$, the inner product of local feature may stress more on base class objects while neglect novel class objects. Thus, we propose the ***inter-image attention*** to alleviate this. At first, the mean descriptor $\mu_Q = \frac{1}{m}\sum_{j=1}^{m} f_\theta(\mathbf{x}_Q)[j]$ is computed. Then, the inter-image attention is calculated as:

$$\mathbf{p}_S^{\text{inter}}[i] = \frac{\exp(\langle f_\theta(\mathbf{x}_S)[i], \mu_Q\rangle^{1/2})}{\sum_{j=1}^{m}\exp(\langle f_\theta(\mathbf{x}_S)[j], \mu_Q\rangle^{1/2})}\,. \tag{4}$$

Finally, the intra-image and inter-image attentions are combined and normalized:

$$\mathbf{p}_S^{\text{attn}}[i] = \frac{\mathbf{p}_S^{\text{intra}}[i] + \mathbf{p}_S^{\text{inter}}[i]}{\sum_{j=1}^{m}(\mathbf{p}_S^{\text{intra}}[j] + \mathbf{p}_S^{\text{inter}}[j])}\,. \tag{5}$$

The query attention $\mathbf{p}_Q^{\text{attn}}$ is calculated in a similar way. Then, $\mathbf{p}_S^{\text{attn}}, \mathbf{p}_Q^{\text{attn}}$ are utilized to replace original $\mathbf{p}_S, \mathbf{p}_Q$ in Eqn. (2), which serve as the bilateral normalization to reduce the weights of background features. This alleviates the object-to-background mismatch issue.

*4) Scale-consistency:* To address the scale mismatch issue, we generate multi-scale local features for the support image while keep query features unchanged. Specifically, multiple pooling and up-sampling operations are applied on $f_\theta(\mathbf{x}_S)$ to get a list of local features: $[f_\theta(\mathbf{x}_S)_1 \in \mathbb{R}^{h_1 w_1 \times c}, \ldots, f_\theta(\mathbf{x}_S)_N \in \mathbb{R}^{h_N w_N \times c}]$. They are concatenated to get the multi-scale support features $f_\theta(\mathbf{x}_S)_{\text{MS}} \in \mathbb{R}^{\Sigma_i h_i w_i \times c}$. The multi-scale design takes several intuitions into account: (1) we use simple pooling instead of convolution or other parameterized modules, which is more efficient and avoids potential overfitting caused by few-shot data; (2) the asymmetric scheme and concatenation make it viable to search for a consistent scale from support features to match the query; (3) even if the object is disproportionately scaled, we can find a combination across multiple scales for a better matching.

Given the bilateral normalization and scale-consistency, we can now update the problem definition in Eqn. (2). We update $\mathbf{C} \in \mathbb{R}^{\Sigma_i h_i w_i \times hw}$ with $\mathbf{C}_{ij} = D(f_\theta(\mathbf{x}_S)_{\text{MS}}[i], f_\theta(\mathbf{x}_Q)[j])$, $\mathbf{p}_S = \mathbf{p}_S^{\text{attn}}$, and $\mathbf{p_Q} = \mathbf{p}_Q^{\text{attn}}$.

**Theorem 2** (Proposition 3.4 in [53]) *In problem (2), for general $\mathbf{p}_S, \mathbf{p}_Q$ (i.e., $\mathbf{p}_S \neq \frac{1}{n}\mathbf{1}, \mathbf{p}_Q \neq \frac{1}{m}\mathbf{1}$), the solution $\mathbf{M}^*$ contains no more than $n+m-1$ nonzero entries, where $n = \sum_i h_i w_i$.*

**Algorithm 1** Sinkhorn algorithm to solve problem (6)

**Input:** $\mathbf{p}_S, \mathbf{p}_Q, \mathbf{C}, \epsilon, t_{\max}$
Initialize $\mathbf{K} = e^{-\mathbf{C}/\epsilon}, \mathbf{b} \leftarrow \mathbf{1}, t \leftarrow 0$
**while** $t \leq t_{\max}$ and not converge **do**
  $\mathbf{a} = \mathbf{p}_S/(\mathbf{K}\mathbf{b})$
  $\mathbf{b} = \mathbf{p}_Q/(\mathbf{K}^\top\mathbf{a})$
**end while**
**Output:** $\mathbf{M}^\epsilon = \text{diag}(\mathbf{a})\mathbf{K}\text{diag}(\mathbf{b})$

From Theorem 2, we can see that with the proposed bilateral normalization and scale-consistent modules, problem (2) still owns good properties to induce a sparse matching matrix $\mathbf{M}^*$. Since $n \neq m$, $\mathbf{M}^*$ can be seen as a relaxed matching matrix. Compared with the original solution, the updated $\mathbf{M}^*$ not only solves the object mismatch by misaligned position, but also addresses the background clutter and scale mismatch with the proposed bilateral normalization and scale-consistency modules. We combine all the proposed modules (optimal local feature matching, bilateral normalization, scale-consistency) in the unified optimal transport problem (Eqn. (2)).

### C. Optimization with Sinkhorn Algorithm

Problem (2) can be solved with network simplex or interior point methods. However, it requires a high complexity of $O(m^3 \log m)$. To reduce the complexity, we solve an entropy-regularized approximation of Eqn. (2):

$$\begin{aligned}\mathbf{M}^\epsilon &= \arg\min_{\mathbf{M}} \sum_{ij} \mathbf{M}_{ij}\mathbf{C}_{ij} + \epsilon H(\mathbf{M})\,,\\ s.t.\quad &\mathbf{M} \geq \mathbf{0}, \mathbf{M}\mathbf{1} = \mathbf{p}_S, \mathbf{M}^\top\mathbf{1} = \mathbf{p}_Q\,,\end{aligned} \tag{6}$$

where $H(\mathbf{M}) = \sum_{ij}\mathbf{M}_{ij}(\log \mathbf{M}_{ij} - 1)$ is the negative entropy and $\epsilon > 0$ is the regularization parameter. Problem (6) is strongly convex and can be solved using Sinkhorn algorithm [51] as shown in Algorithm 1.

**Remark 1** (Approximation Analysis [54]) For the sake of simplicity, we assume $n = m$. The convergence analysis can be performed as follow: by setting $\epsilon = \frac{4\log m}{\tau}$, $\|\mathbf{C}\|_\infty = \max_{ij}|\mathbf{C}_{ij}| \leq L$, Algorithm 1 runs in $O(m^2L^3(\log m)\tau^{-3})$ time to ensure that $\sum \mathbf{M}^\epsilon\mathbf{C} \leq \sum \mathbf{M}^*\mathbf{C} + \tau$. In other words, Eqn. (6) computes a $\tau$-approximation of Eqn. (3) in $O(m^2L^3(\log m)\tau^{-3})$ operations.

*Complexity Analysis.* According to Remark 1, the time complexity of solving Problem (6) with Algorithm 1 is $O(m^2 \log m)$. Calculating the cost matrix $\mathbf{C}$ needs to call $m^2$ times the distance function $D$. So, the overall complexity of our method is $O(m^2 \log m + m^2 c)$.

There are several benefits by introducing entropy regularization. First, the complexity is decreased by an order of magnitude but the solution $\mathbf{M}^\epsilon$ is still near to the original $\mathbf{M}^*$ for small $\epsilon$ according to Remark 1. Second, Algorithm 1 only involves simple matrix operations, which is efficient and differentiable for end-to-end training in CNNs. Third, the regularization has the potential effect of preventing over-aggressive matching, thus improving the generalization ability.

Once $\mathbf{M}^\epsilon$ is obtained, the distance can be computed:

$$\text{Distance}(\mathbf{x}_S, \mathbf{x}_Q) = \sum_{ij}\mathbf{M}_{ij}^\epsilon\mathbf{C}_{ij}\,, \tag{7}$$

which is called **Bilaterally-normalized Scale-consistent Sinkhorn Distance (BSSD)**. The predicted probability for $\mathbf{x}_Q$ is calculated as:

$$p(y_Q = k|\mathbf{x}_Q) = \frac{\exp(\text{Distance}(\mathbf{x}_{S_k}, \mathbf{x}_Q)/\tau)}{\sum_{k'} \exp(\text{Distance}(\mathbf{x}_{S_{k'}}, \mathbf{x}_Q)/\tau)}, \quad (8)$$

where $\tau$ is the temperature similar to recent work [14], [33], [55], $\mathbf{x}_{S_k}$ is the support image of class $k$, the average is applied for more than one images.

## IV. Experiments

In this section, we first describe the benchmark datasets and implementation details. Then ablation study is performed to verify the effectiveness of each proposed component, followed by the qualitative results to provide further intuitions. At last, we compare with the state-of-the-art methods.

### A. Datasets

We perform experiments on three common benchmarks: *mini*ImageNet [6], *tiered*ImageNet [56], and FC100 [57].

***mini*ImageNet** was originally proposed in [6]. It is a subset of ImageNet ILSVRC-2012 and contains $60,000$ images of resolution $84 \times 84$, uniformly distributed over 100 classes. We adopt the common split in [58] to get $64, 16, 20$ classes for training, validation and test, respectively.

***tiered*ImageNet** was proposed in [56] as a large-scale few-shot benchmark. It is also a subset of ImageNet ILSVRC-2012, containing $608$ classes from $34$ super-categories, which are then split to $20, 6, 8$ super-categories, resulting in $351, 97, 160$ classes as training, validation and test set respectively. Similarly, the image size is $84 \times 84$. Since base classes and novel classes come from different super-categories, the semantic gap between training and test phase is larger than *mini*ImageNet.

**FC100** is a few-shot version of CIFAR-100 originally proposed in [57]. Similar to *mini*ImageNet, it has $60,000$ images uniformly distributed over 100 classes. Similar to *tiered*ImageNet, the 100 classes of FC100 are grouped into 20 superclasses, with 12 (60 classes) for training, 4 (20 classes) for validation and 4 (20 classes) for test.

### B. Implementation Details

For a fair comparison with previous works, we consider two commonly used convolutional neural networks as the feature backbone: (1) A 4-layer convolution network (ConvNet) with 64 filters in each layer, following the same architecture in [6], [7], [14] and (2) A 12-layer residual network (ResNet) used in [8], [14]. To get the local feature maps, we remove the last pooling layer of the networks. Concretely, an image of size $84 \times 84$ results in a feature map of $5 \times 5 \times C$[1], *i.e.*, 25 local features. Then, we apply a multi-scale up/down-sampling on the support image to get feature maps of $\{10 \times 10, 5 \times 5, 3 \times 3, 1 \times 1\}$ (except where stated otherwise) and concatenate them to obtain 135 local features.

As a common practice in state-of-the-art literature [14], [33], we apply a feature pre-training step followed by a episodic meta-training step. In pre-training step, we utilize an Adam optimizer with the initial learning rate of $0.002$ for ConvNet. And for ResNet, we use SGD with momentum and set learning rate to $0.001$, momentum to $0.9$. In meta-training step, the learning rate is scaled by a factor $0.1$. In all stages, the weight decay is fixed as $0.0005$. At meta-test time, we follow the new and more trustworthy evaluation setting [14], [27] to randomly sample $10,000$ $N$-way $K$-shot tasks instead of $600$ in previous setting. The average accuracy and 95% confidence interval are reported. The distance metric $D$ is instantiated as $D(\mathbf{u}, \mathbf{v}) = 1 - \mathbf{u}^\top \mathbf{v}/\|\mathbf{u}\|\|\mathbf{v}\|$. The regularization $\epsilon$ is set to $0.05$, max iteration $t_{\max}$ is set to $10$. The softmax temperature $\tau$ in meta-training and meta-test stage is set to $0.02$.

[1] $C$ is feature dimension. C is 64 for ConvNet and 640 for ResNet

### C. Ablation Study

In this subsection, we perform various experiments on *mini*ImageNet to show the effectiveness of our BSSD method.

**Ablation on Sinkhorn distance and feature sizes.** To investigate the effect of Sinkhorn distance on various local/global feature sizes, we re-implement two popular methods: ProtoNet [7] and MatchingNet [6] with various feature map sizes. For local features, we directly flatten the feature map to be a feature vector, which is a straightforward way of applying local features. In addition, we compare with two recent local feature methods, *i.e.*, DN4 [26] and DC [27]. *For our method, we report the plain version here, i.e., only preserving the Sinkhorn distance module and removing both the bilateral normalization and the scale-consistency modules.* This helps to study the pure effect of optimal local feature matching strategy induced by Sinkhorn distance. To ensure a fair comparison, we re-implement all methods to apply a feature pre-training step followed by a meta-training step. All methods utilize the same ConvNet and ResNet architecture, except for the last pooling layer that can generate various feature map sizes. To get a feature map of size $10 \times 10$, we remove the max-pooling accompanying the last convolution layer for ConvNet and ResNet.

The results are shown in Table I. First, simple flattening local features for ProtoNet [7] and MatchingNet [6] does not yield consistent improvements against global features. Due to object mismatch, the simple flattening may introduce inconsistent feature comparison, thus harming the performance. Second, DN4 [26] and Dense [27] are comparable or slightly better than global feature matching of ProtoNet [7] and MatchingNet [7], indicating that their strategies for local feature matching are beneficial but sub-optimal. Third, the proposed BSSD shows consistent and significant improvements against all local and global methods, which corroborates the effectiveness of optimal local feature matching by Sinkhorn distance. Fourth, by increasing feature map size from $5 \times 5$ to $10 \times 10$, DN4, DC and BSSD all gain further improvements for 1-shot setting due to the scarcity of training images. For 5-shot setting with more training images, BSSD continues to increase for larger map size while DN4 and DC do not. We conjecture that with more training images, local features of $10 \times 10$ are nearly saturated and fail to offer additional information.

**Effectiveness of each component.** We then study the benefits of each of the proposed components, including intra-

TABLE I
ABLATION STUDY ON VARIOUS DISTANCE AND FEATURE SIZES. FOR A FAIR COMPARISON, WE RE-IMPLEMENT ALL METHODS WITH THE SAME PRE-TRAINED BACKBONE. IN OUR BSSD METHOD, THE PLAIN VERSION IS REPORTED HERE, *i.e.*, ONLY PRESERVING THE SINKHORN DISTANCE MODULE AND REMOVING BOTH THE BILATERAL NORMALIZATION AND THE SCALE-CONSISTENCY MODULES.

| Method | Distance | Feature | Size | ConvNet | | ResNet | |
|---|---|---|---|---|---|---|---|
| | | | | 1-shot | 5-shot | 1-shot | 5-shot |
| ProtoNet [7] | Euclidean | Global | 1x1 | 52.61 | 71.33 | 62.39 | 80.53 |
| | | Local | 5x5 | 51.97 | 70.60 | 62.05 | 79.76 |
| | | Local | 10x10 | 51.14 | 70.06 | 62.33 | 79.50 |
| MatchingNet [6] | Cosine | Global | 1x1 | 52.87 | 67.49 | 63.87 | 78.72 |
| | | Local | 5x5 | 53.66 | 66.68 | 62.69 | 77.95 |
| | | Local | 10x10 | 53.16 | 67.78 | 63.74 | 77.56 |
| DN4 [26] | Cosine | Local | 5x5 | 53.15 | 70.64 | 63.71 | 78.16 |
| | | Local | 10x10 | 54.09 | 71.60 | 64.13 | 77.55 |
| DC [27] | Euclidean | Local | 5x5 | 53.54 | 71.10 | 63.92 | 80.39 |
| | | Local | 10x10 | 54.39 | 70.77 | 63.95 | 79.81 |
| BSSD (ours) | Sinkhorn | Local | 5x5 | **55.03** | **72.28** | **65.06** | **81.62** |
| | | Local | 10x10 | **55.61** | **72.39** | **65.88** | **81.76** |

TABLE II
EFFECTIVENESS OF THE PROPOSED COMPONENTS ON *mini*IMAGENET.

| Intra-Attention | Inter-Attention | Scale-Consistency | ConvNet | | ResNet | |
|---|---|---|---|---|---|---|
| | | | 1-shot | 5-shot | 1-shot | 5-shot |
| | | | 55.03 | 72.28 | 65.06 | 81.62 |
| ✓ | | | 55.48 | 73.04 | 65.54 | 82.23 |
| | ✓ | | 55.43 | 72.95 | 65.63 | 82.17 |
| ✓ | ✓ | | 55.76 | 73.13 | 65.90 | 82.34 |
| | | ✓ | 55.40 | 72.92 | 66.13 | 82.38 |
| ✓ | | ✓ | 56.09 | 73.03 | 66.39 | 82.69 |
| | ✓ | ✓ | 56.01 | 73.16 | 66.79 | 83.09 |
| ✓ | ✓ | ✓ | **56.53** | **73.44** | **67.28** | **83.48** |

TABLE III
PERFORMANCE WITH VARIOUS FEATURE MAP SIZES ON *mini*IMAGENET.

| Query size | Support size | ConvNet | | ResNet | |
|---|---|---|---|---|---|
| | | 1-shot | 5-shot | 1-shot | 5-shot |
| 5 | 5 | 55.76 | 73.13 | 65.90 | 82.34 |
| 10 | 10 | 56.21 | 73.18 | 66.52 | 82.71 |
| 5,3,1 | 5 | 55.93 | 73.08 | 66.72 | 82.86 |
| 10,5,3,1 | 5 | 56.21 | 73.33 | 66.94 | 83.08 |
| 5 | 5,3,1 | 56.01 | 73.14 | 66.77 | 82.73 |
| 5 | 10,5,3,1 | **56.53** | **73.44** | **67.28** | **83.48** |
| 5,3,1 | 5,3,1 | 56.03 | 73.19 | 66.60 | 83.10 |
| 10,5,3,1 | 10,5,3,1 | 55.98 | 73.30 | 66.47 | 82.74 |

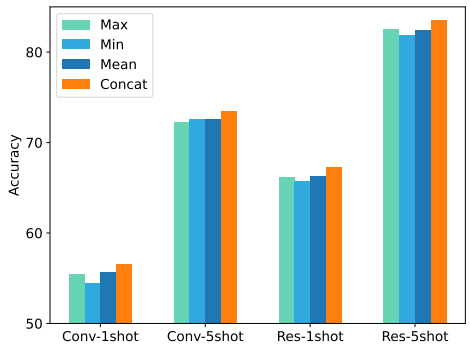



Fig. 4. Accuracy with different multi-scale strategies. The proposed "Concat" achieves the best performance on all settings.

image attention, inter-image attention and scale-consistency. From Table II, we can see that either intra-image attention or inter-image attention alone can obtain accuracy improvements over the baseline model (plain Sinkhorn distance). Combing the two attentions introduces further improvements, showing that they are complementary. The scale-consistency with neither intra-attention nor inter-attention attention can improve the performance of of the plain baseline, corroborating its effectiveness to address the scale mismatch issue. When both attentions and scale-consistency modules are applied, the performance outperforms each single module, which verifies their collaboration. The final model (*i.e.*, BSSD) obtains much better accuracy than plain Sinkhorn distance (+1.5% in most cases) and all other variants. The above analysis indicates that all the proposed components are complementary and indispensable in the proposed BSSD framework.

**Influence of different scales.** In our method, we apply an asymmetric multi-scale design, *i.e.*, keeping query map size as $5 \times 5$ and changing support map size to $\{10 \times 10, 5 \times 5, 3 \times 3, 1 \times 1\}$. This asymmetric multi-scale design can be explained from several perspectives. From an accuracy view, in Table III, the proposed size configuration achieves the best performance among all configurations. This is due to the effective up/down-scaling of the support map, which makes our method possible to capture both the zooming-in and zooming-out relationships between support and query images. Therefore, the scale mismatch issue is well-addressed. From an efficiency view, the proposed size configuration balances the accuracy and efficiency. Our configuration has a complexity of $O(25 \times 135)$ with the quadratic Sinkhorn algorithm, which is relatively small among all sizes. Moreover, the multi-scale operation is applied only on the support maps rather than on the query maps. This choice is derived from a practical consideration. In few-shot setting, the support images are the

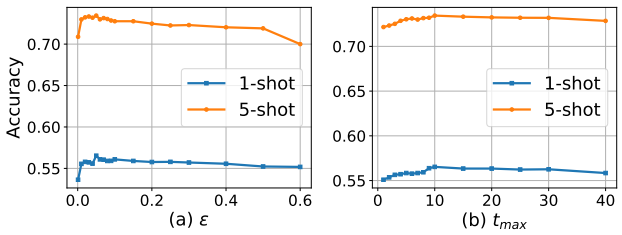



Fig. 5. Accuracy with different values of the parameters $\epsilon$ and $t_{\max}$ using ConvNet on *mini*ImageNet.

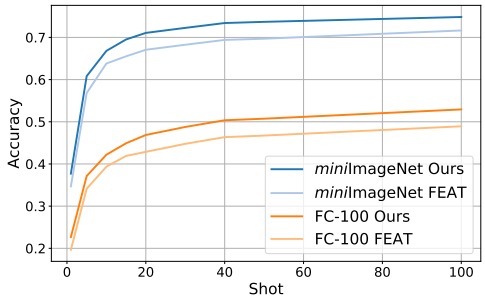



Fig. 6. Performance of various shots on *mini*ImageNet and FC-100.

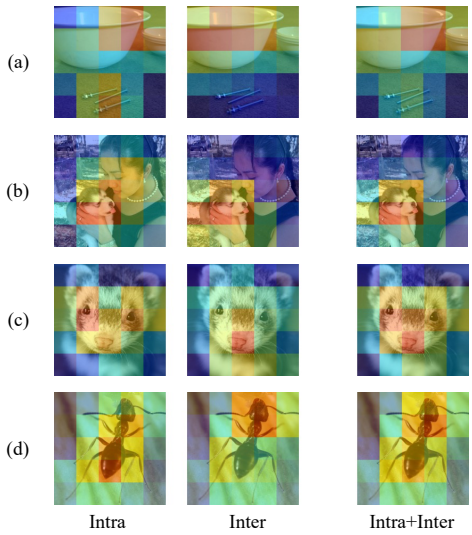



Fig. 7. Visualization of the intra-image and inter-image attention maps.

training set whose number is supposed to be small (*e.g.*, 1 or 5), while the query images are the test set whose number is varying and usually much larger.

**Different scale-consistency strategies.** Given the multi-scale support maps, there are several strategies to apply scale-consistent Sinkhorn distance, such as concatenate features of all scales (*i.e.*, BSSD) or compute distance in each scale and then aggregate with the Max/Min/Mean operation. The results are shown in Fig. 4. The proposed "Concat" achieves the best performance on all settings, corroborating its effectiveness. Since objects can be disproportionately scaled during non-rigid deformation, "Max", "Min" and "Mean" are unable to handle this within the same scale. In contrast, with concatenation, the query image can search for a combination of various scales from the support image, which potentially addresses the non-proportional object scaling issue.

**Influence of the parameters $\epsilon$ and $t_{\max}$.** To investigate the hyper-parameter $\epsilon$, we change it from $0.001$ to $0.5$ for ConvNet on *mini*ImageNet. From Fig. 5(a), we can see that accuracy is quite stable w.r.t $\epsilon$. The best performance is achieved when $\epsilon = 0.05$. The accuracy drops quickly when $\epsilon = 0.001$ as a too small $\epsilon$ leads to numerical issue. Similarly, we vary the parameter $t_{\max}$ from 1 to 40 and report the results in Fig. 5(b). It can be seen that our method is stable w.r.t $t_{\max}$ and the best result is obtained around $t_{\max} = 10$.

**Performance of various shots per class.** To demonstrate the universality of our method, we evaluate the performance of our method and the second best in Table IV (i.e., FEAT [14]) by increasing $K$ from 1 to 100. With the gradual increasing of $K$, the few-shot classification task approaches the general image classification task. From Fig. 6, it is shown that the accuracy acutely increases from 1-shot to 5-shot and then gradually increases until 100-shot. For all different shots, our method surpasses FEAT with a noticeable margin, which indicates the better performance of conducting both the few-shot and general image classification tasks.

### D. Qualitative Results

**Intra-image and inter-image attentions.** The intra-image attention focuses on the objectness of an image, while the inter-image attention considers if the query image contains the same object as the support image. In Fig. 7(a), the intra-image attention map shows high activations for both the bowl and steel nail, but the inter map only focuses on the bowl (*i.e.*, target object). Their combination decreases the non-target nail activations. Similarly, in Fig. 7(b), the non-target human face is also diminished. In some cases, the two attentions are complementary. In Fig. 7(c), the intra map focuses on the eyes area while the inter map focuses on the nose area. Their combination strengthens each individual map by activating both the eyes and nose. Similarly, in Fig. 7(d), the intra map activates the tail part while the inter map activates the head part. Their combination focuses on the whole ant body. The proposed attentions can collaboratively highlight the target objects, then guide the object matching.

**Optimal matching visualization.** In Fig. 8, we show the top 2 matching as well as the corresponding optimal matching matrix $\mathbf{M}$. In Fig. 8(a), the two guitars have similar scales, thus the matching boxes having the same size. In Fig. 8(b) and 8(d), the left and right images have different zoom in/out options, but our algorithm successfully finds the proper matching scales to focus on the same object parts. In Fig. 8(c), the head and leg of the right lion are unevenly scaled due to viewpoint changes. The proposed BSSD handles this issue by seeking for the consistent scales from multiple support scales to match the query scale.

The optimal matching matrix $\mathbf{M}$ sparsely activates the corresponding areas, which is consistent with the Theorems in Section III-B. Moreover, if the objects have similar positions (Fig. 8(a)), the resulting $\mathbf{M}$ shows diagonal activations, otherwise (Fig. 8(b) and 8(d)) $\mathbf{M}$ shows different activations.

TABLE IV
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON *mini*IMAGENET AND *tiered*IMAGENET. WE SHOW THE MEAN ACCURACY AND 95% CONFIDENCE INTERVAL.

| Method | Backbone | *mini*ImageNet | | *tiered*ImageNet | |
|---|---|---|---|---|---|
| | | 1-shot | 5-shot | 1-shot | 5-shot |
| TAPNet [59] | ResNet-12 | 61.65±0.15 | 76.36±0.10 | 63.08±0.15 | 80.26±0.12 |
| DSN [60] | ResNet-12 | 62.64±0.66 | 78.83±0.45 | 66.22±0.75 | 82.79±0.48 |
| Neg-Cosine [61] | ResNet-12 | 63.85±0.81 | 81.57±0.56 | - | - |
| Meta-Baseline [33] | ResNet-12 | 63.17±0.23 | 79.26±0.17 | 68.62±0.27 | 83.29±0.18 |
| TADAM [57] | ResNet-12 | 58.50±0.30 | 76.70±0.30 | - | - |
| MetaOptNet [8] | ResNet-12 | 62.64±0.61 | 78.63±0.46 | 65.99±0.72 | 81.56±0.53 |
| RFS [20] | ResNet-12 | 62.02±0.63 | 79.64±0.44 | 69.74±0.72 | 84.41±0.55 |
| FEAT [14] | ResNet-12 | 66.78±0.20 | 82.05±0.14 | 70.80±0.23 | 84.79±0.16 |
| FBM [62] | ResNet-12 | 61.41±1.87 | 76.11±0.92 | - | - |
| Baseline++ [22] | ResNet-18 | 51.87±0.77 | 75.68±0.63 | - | - |
| Hyperbolic [63] | ResNet-18 | 61.04±0.21 | 77.01±0.15 | - | - |
| CTM [64] | ResNet-18 | 64.12±0.82 | 80.51±0.13 | 68.41±0.39 | 84.28±1.73 |
| SimpleShot [65] | ResNet-18 | 62.85±0.20 | 80.02±0.14 | 69.09±0.22 | 84.58±0.16 |
| TFH-ft [28] | ResNet-18 | 65.07±0.82 | 80.81±0.61 | - | - |
| LEO [66] | WRN-28-10 | 61.76±0.08 | 77.59±0.12 | 66.33±0.05 | 81.44±0.09 |
| PFA [67] | WRN-28-10 | 59.60±0.41 | 73.74±0.19 | - | - |
| Boosting [55] | WRN-28-10 | 64.03±0.46 | 80.68±0.33 | 70.53±0.51 | 84.98±0.36 |
| MixtFSL [68] | WRN-28-10 | 64.31±0.79 | 81.66±0.60 | 68.61±0.91 | 84.08±0.55 |
| BSSD (Ours) | ResNet-12 | **67.28±0.20** | **83.48±0.14** | **71.55±0.23** | **86.13±0.16** |

TABLE V
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON FC100.

| Method | Backbone | 1-shot | 5-shot |
|---|---|---|---|
| SimpleShot [65] | ResNet-10 | 40.13±0.18 | 53.63±0.18 |
| TADAM [57] | ResNet-12 | 40.10±0.40 | 56.10±0.40 |
| MetaOptNet [8] | ResNet-12 | 41.10±0.60 | 55.50±0.60 |
| RFS [20] | ResNet-12 | 42.60±0.70 | 59.10±0.60 |
| DC [27] | ResNet-12 | 42.04±0.17 | 57.05±0.16 |
| MTL [23] | ResNet-12 | 45.10±1.80 | 57.60±0.90 |
| MixtFSL [68] | ResNet-12 | 44.89±0.63 | 60.70±0.67 |
| BSSD (Ours) | ResNet-12 | **47.13±0.26** | **63.59±0.25** |

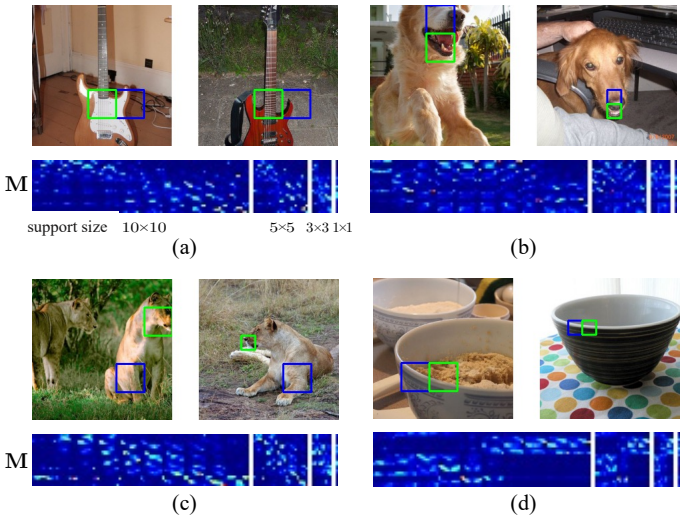



Fig. 8. Visualization of the top matching (same box colors denote a matching pair) and the optimal matching matrix $\mathbf{M}$.

### *E. Comparison with the State of the Art*

Finally, we compare the proposed BSSD with the recent state-of-the-art methods. We report the 5-way 1-shot and 5-way 5-shot results on 3 popular benchmarks: *mini*ImageNet, *tiered*ImageNet, and FC100. As shown in Table IV and Table V, our algorithm outperforms the state-of-the-art methods on all three benchmarks. Among the baselines, TFH-ft [28] generate tensor features and DC [27] applied dense classifiers on each local feature. Since none of them addresses the mismatch issues, they obtain inferior accuracy. For the backbone, although we employ a small network (*i.e.*, ResNet 12), our algorithm outperforms the methods with a larger backbone structure such as ResNet-18 or WideResNet-28.

## V. CONCLUSION

In this paper, we propose the Bilaterally-normalized Scale-consistent Sinkhorn Distance (BSSD) to deal with the few-shot classification problem. Specifically, we utilize the more discriminative local feature maps to replace the widely-used global feature vectors, leading to more consistent and transferable features. Direct local feature matching encounters two practical issues: object mismatch and scale mismatch, which are caused by misaligned position, cluttered background, and inconsistent object scales. To address these issues, we propose three novel modules: optimal local feature matching, bilateral normalization, and scale-consistency, which are incorporated in a unified end-to-end BSSD framework. Extensive experiments show the effectiveness of each designed module, and we achieve the state-of-the-art on three benchmark datasets.

## REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[4] S. Tsutsui, Y. Fu, and D. Crandall, "Meta-reinforced synthetic data for one-shot fine-grained visual recognition," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[5] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, no. 33, 2011.

[6] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[8] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 657–10 665.

[9] R. Hou, H. Chang, M. Bingpeng, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *Advances in Neural Information Processing Systems*, 2019, pp. 4003–4014.

[10] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, and L. Sigal, "Multi-level semantic feature augmentation for one-shot learning," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4594–4605, 2019.

[11] C. Liu, C. Xu, Y. Wang, L. Zhang, and Y. Fu, "An embarrassingly simple baseline to one-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 922–923.

[12] X. Li, Q. Sun, Y. Liu, Q. Zhou, S. Zheng, T.-S. Chua, and B. Schiele, "Learning to self-train for semi-supervised few-shot classification," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 276–10 286.

[13] H.-J. Ye, H. Hu, and D.-C. Zhan, "Learning adaptive classifiers synthesis for generalized few-shot learning," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1930–1953, 2021.

[14] 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*, 2020, pp. 8808–8817.

[15] 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-Volume 70*, 2017, pp. 1126–1135.

[16] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[17] H.-J. Ye, X.-R. Sheng, and D.-C. Zhan, "Few-shot learning with adaptively initialized task optimizer: a practical meta-learning approach," *Machine Learning*, vol. 109, no. 3, pp. 643–664, 2020.

[18] N. Lai, M. Kan, C. Han, X. Song, and S. Shan, "Learning to learn adaptive classifier–predictor for few-shot learning," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 8, pp. 3458–3470, 2020.

[19] R.-Q. Wang, X.-Y. Zhang, and C.-L. Liu, "Meta-prototypical learning for domain-agnostic few-shot recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[20] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: a good embedding is all you need?" in *European conference on computer vision*. Springer, 2020, pp. 266–282.

[21] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *International Conference on Learning Representations*, 2019.

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

[23] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 403–412.

[24] X. Yu, Y. Tian, F. Porikli, R. Hartley, H. Li, H. Heijnen, and V. Balntas, "Unsupervised extraction of local image descriptors via relative distance ranking loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[25] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, "Sosnet: Second order similarity regularization for local descriptor learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 016–11 025.

[26] 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*, 2019, pp. 7260–7268.

[27] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, "Dense classification and implanting for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9258–9267.

[28] M. Lazarou, T. Stathaki, and Y. Avrithis, "Tensor feature hallucination for few-shot learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3500–3510.

[29] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies," *Frontiers of Information Technology & Electronic Engineering*, vol. 22, no. 12, pp. 1551–1558, 2021.

[30] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 528–537.

[31] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[32] 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*, 2018, pp. 1199–1208.

[33] Y. Chen, X. Wang, Z. Liu, H. Xu, and T. Darrell, "A new meta-baseline for few-shot learning," *arXiv preprint arXiv:2003.04390*, 2020.

[34] R.-Q. Wang, X.-Y. Zhang, and C.-L. Liu, "Meta-prototypical learning for domain-agnostic few-shot recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6990–6996, 2021.

[35] Y. Liu, J. Lee, L. Zhu, L. Chen, H. Shi, and Y. Yang, "A multi-mode modulator for multi-domain few-shot classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8453–8462.

[36] Z. Ji, Z. Hou, X. Liu, Y. Pang, and J. Han, "Information symmetry matters: a modal-alternating propagation network for few-shot learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 1520–1531, 2022.

[37] X. Liu, Z. Ji, Y. Pang, J. Han, and X. Li, "Dgig-net: Dynamic graph-in-graph networks for few-shot human–object interaction," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 7852–7864, 2021.

[38] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, and E. Ricci, "A unified objective for novel class discovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9284–9292.

[39] Z. Zhong, L. Zhu, Z. Luo, S. Li, Y. Yang, and N. Sebe, "Openmix: Reviving known knowledge for discovering novel visual categories in an open world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9462–9470.

[40] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "Vehiclenet: Learning robust visual representation for vehicle re-identification," *IEEE Transactions on Multimedia*, vol. 23, pp. 2683–2693, 2020.

[41] X. Wang, L. Zhu, Y. Wu, and Y. Yang, "Symbiotic attention for egocentric action recognition with object-centric alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[43] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8460–8469.

[44] C. Wang, J. Xiao, Y. Han, Q. Yang, S. Song, and G. Huang, "Towards learning spatially discriminative feature representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1326–1335.

[45] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 990–11 997.

[46] R. Xu, P. Liu, L. Wang, C. Chen, and J. Wang, "Reliable weighted optimal transport for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4394–4403.

[47] T. Long, Y. Sun, J. Gao, Y. Hu, and B. Yin, "Domain adaptation as optimal transport on grassmann manifolds," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[48] C. Bunne, D. Alvarez-Melis, A. Krause, and S. Jegelka, "Learning generative models across incomparable spaces," in *International Conference on Machine Learning*. PMLR, 2019, pp. 851–861.

[49] H. Xu, D. Luo, H. Zha, and L. C. Duke, "Gromov-wasserstein learning for graph matching and node embedding," in *International conference on machine learning*. PMLR, 2019, pp. 6932–6941.

[50] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 5, pp. 1592–1601, 2019.
[51] M. Cuturi, "Sinkhorn distances: lightspeed computation of optimal transport." in *NIPS*, vol. 2, no. 3, 2013, p. 4.
[52] T. Wang, Z. Wu, and D. Wang, "Visual perception generalization for vision-and-language navigation via meta-learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
[53] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.
[54] J. Altschuler, J. Niles-Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via sinkhorn iteration," *Advances in neural information processing systems*, vol. 30, 2017.
[55] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8059–8068.
[56] 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 *International Conference on Learning Representations*, 2018.
[57] B. N. Oreshkin, P. R. López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *NeurIPS*, 2018.
[58] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *International Conference on Learning Representations*, 2017.
[59] S. W. Yoon, J. Seo, and J. Moon, "Tapnet: Neural network augmented with task-adaptive projection for few-shot learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7115–7123.
[60] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4136–4145.
[61] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 438–455.
[62] P. Yang, S. Ren, Y. Zhao, and P. Li, "Calibrating cnns for few-shot meta learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2090–2099.
[63] P. Fang, M. Harandi, and L. Petersson, "Kernel methods in hyperbolic spaces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 665–10 674.
[64] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1–10.
[65] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "Simpleshot: Revisiting nearest-neighbor classification for few-shot learning," *arXiv preprint arXiv:1911.04623*, 2019.
[66] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *International Conference on Learning Representations*, 2018.
[67] S. Qiao, C. Liu, W. Shen, and A. L. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7229–7238.
[68] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Mixture-based feature space learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9041–9051.

**Yanbin Liu** received the B.E. degree and M.S. degrees from Tianjin University, China, in 2013 and 2015, respectively. He received the Ph.D. degree from the University of Technology Sydney, Australia, in 2021. He is currently a Research Fellow with the School of Computing, Australian National University. His research interests lie in machine learning and deep learning for computer vision problems, especially learning with limited labeled data.

**Linchao Zhu** (Member, IEEE) received the B.E. degree from Zhejiang University, China, in 2015, and the Ph.D. degree in computer science from the University of Technology Sydney, Australia, in 2019. He is a Research Professor with the College of Computer Science and Technology, Zhejiang University, China. His research interests are video analysis and understanding.

**Xiaohan Wang** received the Ph.D. degree in computer science from University of Technology Sydney, Australia, in 2021. He received the B.E. degree from University of Science and Technology of China, China, in 2017. He is currently a postdoctoral researcher with the College of Computer Science and Technology, Zhejiang University, China. His research interest includes video analysis, egocentric vision and multi-modal understanding.

**Makoto Yamada** received the PhD degree in statistical science from The Graduate University for Advanced Studies (SOKENDAI, The Institute of Statistical Mathematics), Tokyo, in 2010. He has held positions as a postdoctoral fellow with the Tokyo Institute of Technology from 2010 to 2012, as a research associate with NTT Communication Science Laboratories from 2012 to 2013, as a research scientist with Yahoo Labs from 2013 to 2015, as an assistant professor with Kyoto University from 2015 to 2017, as a team leader with RIKEN from 2017 to 2023, and as an associate professor with Kyoto University from 2018 to 2023. Currently, he is an associate professor at the Okinawa Institute of Science and Technology (OIST) His research interests include machine learning and its application to biology, natural language processing, and computer vision. He published more than 50 research papers in premium conferences and journals such as NeurIPS, AISTATS, ICML, AAAI, IJCAI, and TPAMI, and won the WSDM 2016 Best Paper Award.

**Yi Yang** (Senior Member, IEEE) received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He was a Professor with the University of Technology Sydney, Ultimo, NSW, Australia. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia analysis and video semantics understanding.