EASE: Unsupervised Discriminant Subspace Learning for Transductive Few-Shot Learning

Hao Zhu[†,§], Piotr Koniusz[*,§,†]
[†]The Australian National University [§]Data61/CSIRO
allenhaozhu@gmail.com, firstname.lastname@anu.edu.au

Abstract

Few-shot learning (FSL) has received a lot of attention due to its remarkable ability to adapt to novel classes. Although many techniques have been proposed for FSL, they mostly focus on improving FSL backbones. Some works also focus on learning on top of the features generated by these backbones to adapt them to novel classes. We present an unsupErvised discriminAnt Subspace lEarning (EASE) that improves transductive few-shot learning performance by learning a linear projection onto a subspace built from features of the support set and the unlabeled query set in the test time. Specifically, based on the support set and the unlabeled query set, we generate the similarity matrix and the dissimilarity matrix based on the structure prior for the proposed EASE method, which is efficiently solved with SVD. We also introduce conStraIned wAsserstein MEan Shift clustEring (SIAMESE) which extends Sinkhorn K-means by incorporating labeled support samples. SIAMESE works on the features obtained from EASE to estimate class centers and query predictions. On the mini-ImageNet, tiered-ImageNet, CIFAR-FS, CUB and OpenMIC benchmarks, both steps significantly boost the performance in transductive FSL and semi-supervised FSL.

1. Introduction

Supervised end-to-end learning has been extremely successful in computer vision, speech, and machine translation tasks due to larger datasets and the rapid development of deep convolutional architectures. However, with the current learning paradigm, limited data is an obstacle in training a sufficiently good model for inference. In many tasks, annotation is likely to be scarce or costly to obtain, as the annotation process may require expert knowledge (*e.g.*, association of medical images with diseases).

In contrast to the requirement of big data in deep learning, humans learn new objects from a few examples. Inspired by the ability of biological vision, researchers proposed few-shot learning (FSL) [5]. FSL algorithms can be broadly classified into three categories: metric learning, meta-learning, and transfer learning. The goal of metric-based learning approaches is to learn a mapping from images to an embedding space in which images of the same class are closer to each other, whereas images of different classes are separated apart. We expect this property to apply to classes that have not been seen before. Meta-learning is tasked with resolving an individual task-specific optimization that can be easily adapted to new tasks without overfitting. Transfer learning includes pre-training a feature extractor in the first stage and then learning to reuse this knowledge to obtain a classifier on new samples.

Several recent studies [4, 9, 10, 15, 26, 31] explored transductive inference for few-shot tasks. At the test time, transductive few-shot methods perform the class label inference jointly for all the unlabeled query samples of the task, rather than one sample at a time, as in the case of inductive inference [17, 25, 38, 40, 52–55]. Therefore, transductive few-shot methods typically perform better than their inductive counterparts. The core idea of transductive FSL methods is to connect the query set and the support set with a pseudo-label or via a similarity measure *e.g.*, the Gaussian kernel. However, such methods ignore the potential structure among the data points in the support set and the query set. In this paper, we argue that features in the inference step can be approximately drawn from a union of multiple subspaces, and thus the sample affinity matrix follows the block-diagonal prior. Based on this prior, we generate the similarity matrix for all samples in each task based on subspace clustering with a low-rank representation, instead of measuring the Euclidean distance between samples in the support or query set. By constructing a dissimilarity matrix with a simple assumption, we learn a linear projection for maximizing the similarity and minimizing the dissimilarity via a closed-form solution. Furthermore, we improve the final performance by refining each class mean with unlabeled

[*]The corresponding author. Code: `https://github.com/allenhaozhu/EASE`

data. In experiments, our model outperforms state-of-the-art methods by significant margins, consistently providing improvements across different settings, datasets, and training models. Furthermore, our transductive inference is very fast, with runtimes that are close to the runtimes of inductive inference, and can be used for the large-scale task.

Our contributions are as follows:

i. We propose an assumption that the features in the inference step can be approximately drawn from a union of multiple subspaces, and thus its similarity matrix follows the block-diagonal prior.

ii. We propose an unsupErvised discriminAnt Subspace lEarning (EASE), which is able to learn a discriminant subspace by maximizing the inter-class distance and minimizing the intra-class distance, akin to positive and negative sampling in graph node embedding.

iii. As a minor contribution, we propose conStraIned wAsserstein MEan Shift clustEring (SIAMESE) which extends Sinkhorn K-means by incorporating labels of the support set. SIAMESE works on the features from EASE to estimate class centers and query predictions.

2. Related Work

Few-shot classification. Most few-shot methods are based on the meta-learning paradigm [32, 40, 43], and they rely on episodic learning to perform training and testing. The goal is to learn a model that can efficiently adapt to new tasks with novel categories given a few labeled samples. Approaches [2, 44] demonstrate that meta-training is not required for learning good features for few-shot learning. Instead, they train a typical classification network with two parts: the feature extractor and the classification head. Subsequently, they learn a base network by standard training using the base class data, and use it with a nearest neighbor classifier for samples of novel classes. In this paper, our attention is not on how to learn a better backbone from the base class samples but on how to design the inference stage and improve its performance in transductive and semi-supervised settings given an existing backbone.

Metric-based FSL. Commonly trained using episodes [26, 40], metric learning approaches are characterized by a similarity classifier learnt over a feature space. They focus on learning high-quality and transferable features with a neural network backbone common across tasks. For instance, the matching network [43] uses an end-to-end trainable k-nearest neighbors algorithm on the learnt embedding of the few labeled examples (support set) to predict the classes of the unlabeled samples (query set), whereas the prototypical network [40] further builds a pre-class prototype representation. More recently, Sung *et al*. presented the relation network [41] which learns a nonlinear distance metric via a simple neural network instead of using a fixed linear distance metric *e.g*., Cosine or Euclidean. These methods use mini-batches of episodes to train an end-to-end network, assuming that the training features will be representative of novel test classes. Our motivation is different from such methods in that we employ metric learning on the feature space in the test stage (novel classes) to find a discriminant subspace for classification. Moreover, our approach is unsupervised. This makes it work well even in the 1-shot setting.

Graph-based FSL. One can consider this family as a special case of metric learning because most of the graph-based FSL methods form a graph via an RBF-based adjacency matrix used in the propagation of labels or features. Satorras *et al*. [37] propagate labels by building an affinity matrix between the support set and the unlabeled data. wDAE-GNN [7] generates classification weights with a graph neural network (GNN) and applies a denoising AutoEncoder (DAE) to regularize the representation. Embedding Propagation [34] not only propagates labels but also considers propagating the embedding to decrease the intra-class distance. Set-to-set functions have also been used for the adaptation of embedding [49], where GCN is also used to form an instance of set-to-set functions. However, different from GNN-based methods, we do not use a graph or any affinity matrix to propagate labels or features. Our approach is to estimate the similarity matrix (graph) for given samples based on the block-diagonal prior. Specifically, we employ a self-representation with the low-rank constraint for that. We learn a linear projection based on the graph embedding framework with the similarity (and dissimilarity) matrices, and then project features to a subspace to improve FSL performance in the transductive setting.

Transductive Few-shot Learning and Semi-Supervised Few-Shot Learning. The most common setting in FSL is the inductive setting. In such a scenario, only samples in the support set can be used to fine-tune the model or learn a function for the inference of query labels. In contrast, in the transductive scenario, the model has access to all the query data (unlabeled) that needs to be classified. In [4], the query samples are used in fine-tuning in conjunction with the entropy minimization in order to maximize the certainty of their predictions. Label propagation [26] and embedding propagation are also used in representation learning, as in meta-learning [34]. Lichtenstein *et al*. [23] note that the model pre-trained on base classes cannot represent the novel class very effectively. For this reason, they use PCA or ICA to extract better quality features for the further classification task. Despite PCA and ICA are able to improve the classification performance, they are not designed to act in a discriminant manner via a discriminant criterion step.

In semi-supervised FSL [22, 26, 33, 39], the unlabeled data is provided in addition to the support set and is assumed to have a similar distribution to the target classes

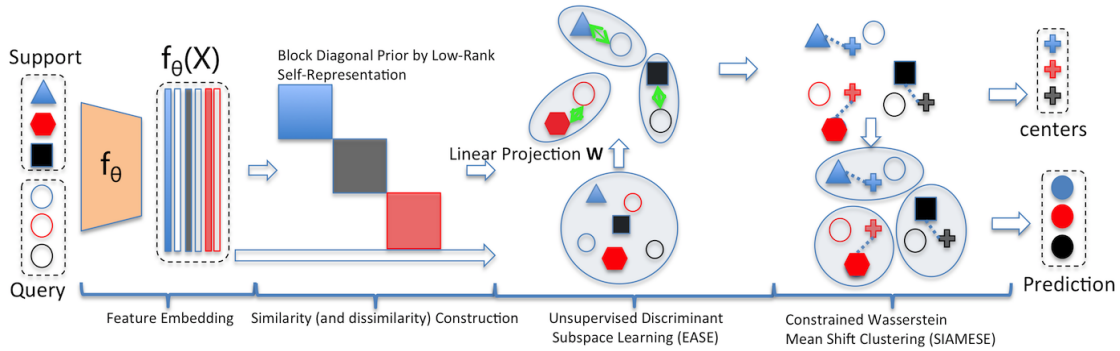

Figure 1. Illustration of our method for transductive few-shot learning. Image features are extracted with a pre-trained CNN. Then these features are used to build similarity and dissimilarity (adjacency) matrices. Based on these matrices, we learn a linear projection by minimizing our EASE objective to take into account the structure prior. SIAMESE works on the features obtained from the EASE projection step to estimate class centers and the predictions of queries. SIAMESE uses labeled samples with labels and unlabeled samples.

(although some unrelated noise samples may be also included). In the LST [22], self-labeling and soft-attention mechanisms are used on the unlabeled samples intermittently with fine-tuning on the labeled and self-labeled data. Similarly to LST, Ren *et al*. [33] update the class prototypes using K-means iterations initialized by the PN prototypes. Their method also includes down-weighting the potential distractor samples (likely not to belong to the target classes) in the unlabeled data. Simon *et al*. [39] use unlabeled examples through the soft-label propagation. Saito *et al*. [36] study the problem of semi-supervised few-shot domain adaptation. In [15, 26, 37], graph neural networks are used for sharing the information between labeled and unlabeled samples in the semi-supervised FSL setting [26, 37]. In [26] a graph construction network is employed to predict the task-specific graph for propagating labels across samples of semi-supervised FSL task. Liu *et al*. [25] argue that there exists a bias between the prototype representations and the expected representations, and provide a simple strategy to rectify the bias based on intra-class and inter-class assumptions. In this paper, we propose an unsupervised metric learning method for few-shot learning to learn a discriminant subspace which can pull similar features closer while pushing dissimilar features further away from each other in test time. Our framework is compatible with transductive and semi-supervised few-shot learning.

3. Methodology

In this section, we introduce our unsupErvised discriminAnt Subspace lEarning (EASE) method and explain its details. Firstly, we define our model with the classifier in prototypical networks [40]. Secondly, we present how to learn a discriminant subspace given similarity and dissimilarity matrices, and we explain our assumptions about the dissimilarity matrix. Thirdly, we demonstrate how to use the so-called self-representation with the low-rank regularization to construct the similarity matrix with the structure prior. Given a task T and features from a feature extractor, we apply subspace clustering via low-rank representation to obtain a similarity matrix. Finally, we reformulate the estimation of class centers and the query prediction as a conStraIned wAsserstein MEan Shift clustEring (SIAMESE), which extends Sinkhorn K-means [11, 12] by using labeled samples with labels and unlabeled samples, and works on the features from EASE. Figure 1 illustrates our method.

3.1. Model Description

In few-shot classification, we are given a small support set of K classes with N labeled examples per class, $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$ where each $\mathbf{x}_i$ is a raw image and $y_i \in \{1, \cdots, K\}$ is the corresponding label. Moreover, $S_k \subset S$ is the set of examples labeled with class k. Prototypical networks [40] compute an M-dimensional representation $\tilde{\boldsymbol{c}}_k \in \mathbb{R}^O$, or prototype, of each class through an embedding function f_θ with learnable parameters θ. Each prototype is the mean vector of the embedded support points belonging to its class:

$$\tilde{\boldsymbol{c}}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\theta(\mathbf{x}_i). \tag{1}$$

Given a distance function $d : \mathbb{R}^O \times \mathbb{R}^O \to \mathbb{R}^+$, prototypical networks use Nearest Class Mean (NCM) classifier to predict the label of a query point $\mathbf{x}$ over distances to the prototypes in the embedding space:

$$k^* = \arg\min_k d\left(f_\theta(\mathbf{x}), \tilde{\boldsymbol{c}}_k\right). \tag{2}$$

However, at the test time, it is hard to update parameters θ of the backbone reliably with few test samples. Thus, given task T, one can define a linear projection $\mathbf{W} \in \mathbb{R}^{O \times O'}$ ($O' \ll O$) and insert it into Eq. (2), leading to:

$$k^* = \arg\min_k d\left(\mathbf{W} f_\theta(\mathbf{x}), \mathbf{W}\tilde{\boldsymbol{c}}_k\right). \tag{3}$$

where $\mathbf{W}$ is a learnable orthonormal matrix learnt during the test time for the given task T. To learn $\mathbf{W}$, many metric learning approaches use relative or absolute similarity constraints *i.e.*, pairwise or triplet-based approaches. The triplet of images can be defined as $(f_\theta(\mathbf{x}_i), f_\theta(\mathbf{x}_j), f_\theta(\mathbf{x}_{j'}))$, where $f_\theta(\mathbf{x}_i)$, $f_\theta(\mathbf{x}_j)$ and $f_\theta(\mathbf{x}_{j'})$ correspond to the feature vectors of an anchor $\mathbf{x}_i$, positive $\mathbf{x}_j$, and negative image $\mathbf{x}_{j'}$, respectively. Moreover, $\mathbf{x}_i$ and $\mathbf{x}_j$ share similar class labels, whereas $x_{j'}$ should have a different class label than the class label of the anchor. A pair of image features corresponding to an image pair $(\mathbf{x}_i, \mathbf{x}_j)$ is denoted as $(f_\theta(\mathbf{x}_i), f_\theta(\mathbf{x}_j))$. It refers to as a positive pair if both images share the same label, or negative pair otherwise. However, the triplet-based loss also suffers from the low number of samples in the test stage. For the 5-way 1-shot setting, it is hard to minimize the intra-class distance because only one sample per class is available. Maximizing the inter-class similarity alone cannot solve the issue that the intra-class distances are often larger than the inter-class ones. Thus, employing the regular metric learning for few-shot learning seems limited because one has not enough positive pairs at the test stage in an episode.

In the following section, we present a novel unsupervised discriminant subspace learning for transductive few-shot learning, which does not require any label information. As we have available only few samples with labels in the support set for each task, we employ the query set which has many samples (to use without labels). Given an insufficient number of labeled samples per episode, we instead transform the metric learning problem into a graph embedding problem and then use similarity and dissimilarity measures as surrogates for the label information.

3.2. Unsupervised Discriminant Subspace Learning

Below, we present our unsupErvised discriminAnt Subspace lEarning (EASE). One can learn a linear projection function for Eq. (3) by maximizing the similarity between similar features and minimizing the dissimilarity between dissimilar features. Given the mapping f_θ, the labeled support set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$ and the query set $Q = \{\mathbf{x}_{L+i}\}_{i=1}^{U}$, we combine $\mathbf{x}_i$ in the support set and the query set into $\mathbf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{L+U}]^\top$ and define EASE as:

$$\mathbf{W}^* = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\min}\, \mathbf{W} f_\theta(\mathbf{X})^\top (\mathbf{L}_{sim} - \mathbf{L}_{dis}) f_\theta(\mathbf{X}) \mathbf{W}^\top, \quad (4)$$

where $\mathbf{L}_{sim} = \mathbf{I} - \tilde{\mathbf{A}}_{sim}$ and $\mathbf{L}_{dis} = \mathbf{I} - \tilde{\mathbf{A}}_{dis}$. Let $\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \tilde{\mathbf{A}}$ and $\mathbf{D} = \mathrm{diag}(d_1, \cdots, d_{L+N})$, where $d_i = \sum_j \mathbf{A}_{ij}$. We explain how we obtain $\mathbf{A}_{sim}$ and $\mathbf{A}_{dis}$ latter in the text. As $\mathbf{L}_{sim} - \mathbf{L}_{dis} = \tilde{\mathbf{A}}_{dis} - \tilde{\mathbf{A}}_{sim}$, we obtain:

$$\mathbf{W}^* = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\min}\, \mathbf{W} f_\theta(\mathbf{X})^\top (\tilde{\mathbf{A}}_{dis} - \tilde{\mathbf{A}}_{sim}) f_\theta(\mathbf{X}) \mathbf{W}^\top, \quad (5)$$

where $\mathbf{A}_{sim}$ and $\mathbf{A}_{dis}$ are two different measurements with the opposite effect. Thus, we introduce parameter $\alpha > 0$ to balance the impact of these both terms, and we obtain:

$$\mathbf{W}^* = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}}{\arg\min}\, \mathbf{W} f_\theta(\mathbf{X})^\top (\alpha\tilde{\mathbf{A}}_{dis} - \tilde{\mathbf{A}}_{sim}) f_\theta(\mathbf{X}) \mathbf{W}^\top. \quad (6)$$

The optimal $\mathbf{W}^*$ can be found by selecting top-k (not bottom-k) eigenvectors of $f_\theta(\mathbf{X})^\top(\tilde{\mathbf{A}}_{sim} - \alpha\tilde{\mathbf{A}}_{dis}) f_\theta(\mathbf{X})$ (the so-called generalised eigenvalue problem).

Dissimilarity Matrix. Although one might design a linear projection based on the similarity relationship alone, we use both the dissimilarity information and the similarity matrix for learning a linear projection. Intuitively, given a K-way N-shot task with B queries for each class, we are targeting an $(N + B) \times K$-way 1-shot problem where off-diagonal entries can be assumed to represent differing entities (on-diagonals represent the same entity). Thus, we form a dissimilarity matrix as the adjacency matrix of a densely connected graph:

$$\mathbf{A}_{dis} = \frac{1}{(N+B)K}\mathbf{e}\mathbf{e}^\top - \mathbf{I}, \quad (7)$$

where $\mathbf{e}$ is an $((N + B)\,K)$-dimensional all-ones vector and $\mathbf{I}$ is the identity matrix. We will discuss the relation between the dissimilarity matrix and PCA later in the text.

Similarity Matrix. To measure the similarity between the pairs of samples, one has to choose a distance (or similarity measure) that will perform well in the FSL setting.

The typical choice for the measure of similarity is the RBF function $Z_{ij} = \exp(-\|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2^2/\sigma)$, $\sigma > 0$ but the RBF function alone does not capture the structure of data. In this paper, we claim that for the K-way few-shot learning task, the expected similarity matrix should be a K-block diagonal matrix, as shown in Figure 1. However, the similarity matrix based on the RBF kernel has no block-diagonal structure.

Low-Rank Representation (LRR) [24] expresses each data point $\mathbf{x}_i$ as a linear combination of other points, $\mathbf{x}_i = \sum_{j\neq i} Z_{ij}\mathbf{x}_j$, and uses the representational coefficient $(|Z_{ij}| + |Z_{ji}|)/2$ to measure the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. LRR takes the correlation structure of data into account, and finds a low-rank representation instead of a sparse representation. In this paper, the LRR is applied in the following rank minimization problem:

$$\underset{\mathbf{Z}}{\arg\min}\, \|f_\theta(\mathbf{X}) - f_\theta(\mathbf{X})\mathbf{Z}\|_F^2 \quad \text{s.t. } \mathrm{rank}(\mathbf{Z}) = K. \quad (8)$$

Eq. (8) is solved in two stages: 1) $\mathbf{Z} = \mathbf{V}^\top\mathbf{V}$, where $\mathbf{V}$ is obtained from the skinny SVD of $f_\theta(\mathbf{X}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, and 2) for each row of $\mathbf{V}$, one only keeps top-K absolute largest entries of $\mathbf{\Sigma}$. Given the feature matrix $f_\theta(\mathbf{X})$, we obtain the representation matrix $\mathbf{Z}$ by solving Eq. (8). The similarity matrix is defined as $\mathbf{W}_{sim} = |\mathbf{Z}| - \mathrm{diag}(|\mathbf{Z}|)$.

Relation to PCA. By PCA, one seeks projection directions with maximal variances akin to our problem below:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}^\top\mathbf{w}=1} \mathbf{w} f_\theta(\mathbf{X})^\top \left(\mathbf{I} - \frac{1}{(N+B)K}\mathbf{e}\mathbf{e}^\top\right) f_\theta(\mathbf{X})\mathbf{w}^\top. \tag{9}$$

We note that $f_\theta(\mathbf{X})^\top \mathbf{A}_{dis} f_\theta(\mathbf{X})$ is the negative covariance matrix for the given $f_\theta(\mathbf{X})$. As each row or column of $\mathbf{A}_{dis}$ are $\mathbf{e}$ or $\mathbf{e}^\top$, we have $\mathbf{D}^{-1/2}\mathbf{A}_{dis}\mathbf{D}^{-1/2} = \frac{1}{(N+B)K}\mathbf{A}_{dis}$.

Theoretical Interpretation. Eq. (6) can be regarded as working with two different loss functions, one preserving the similarity and the other preserving the maximal variance. This is akin to positive and negative sampling strategies in graph node embedding [28,47,56,57], the latter strategy resembles the so-called negative dense graph *e.g.*, formed by uniform node sampling on link nodes.

3.3. conStraIned wAsserstein MEan Shift clustEring (SIAMESE)

In the case of inductive FSL, the prediction is performed independently on each sample, and thus the mean vector is only dependent on the support set of N labeled examples, as shown in Eq. (1), and is fixed when for the given embedded features. However, in the case of transductive FSL, the prediction is performed inclusive of all queries $Q = \{\mathbf{x}_{L+i}\}_{i=1}^{U}$ together. Considering the support set alone to estimate the class mean does not take full advantage of the queries [58]. We embed all examples from S and Q into $\boldsymbol{H} = f_\theta(\mathbf{X})\mathbf{W}^\top$ where $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_{L+U}]^\top \in \mathbb{R}^{(L+U)\times O'}$. One may employ the query set by minimizing:

$$\min_{\mathbf{P},\tilde{\mathbf{C}}} \sum_{i,k} P_{i,k}\|\boldsymbol{h}_i - \tilde{\boldsymbol{c}}_k\|_2^2, \text{ s.t. } \mathbf{P}\cdot\mathbf{1}=\mathbf{1} \text{ and } P_{i,k} \geq 0, \tag{10}$$

by alternating between $\boldsymbol{P} \in \mathbb{R}^{(L+U)\times K}$ and $\tilde{\boldsymbol{C}} \in \mathbb{R}^{K\times O'}$, where $\tilde{\mathbf{C}} = [\tilde{\boldsymbol{c}}_1, ..., \tilde{\boldsymbol{c}}_K]^\top$ is estimated by weighted-averaging the support and query sets with their allocated portions for class k, while P_{ik} is the weight of the i-th sample w.r.t. the k-th center $\tilde{\boldsymbol{c}}_k$ (think the conditional probability $p(y_i = k|\boldsymbol{h}_i)$). We notice that one could use the mean shift algorithm [23] to estimate P_{ik} and then update centers $\tilde{\mathbf{C}}$.

However, Eq. (10) ignores another constraint $\mathbf{1}^\top\mathbf{P} = \mathbf{1}$ which balances the class distribution among K classes [20]. Thus, we reformulate Eq. (10) with the constraint as minimizing the optimal transport distance between $\boldsymbol{H}$ and $\tilde{\boldsymbol{C}}$ under fixed $\tilde{\boldsymbol{C}}$. Given a cost matrix $\boldsymbol{M} \in \mathbb{R}^{(L+U)\times K}$ and $\boldsymbol{M}_{ik} = \|\boldsymbol{h}_i - \tilde{\boldsymbol{c}}_k\|_2^2$, the cost of mapping $\boldsymbol{r} = \mathbf{1}_{L+U}$ to $\boldsymbol{c} = (N+B)\mathbf{1}_K$ using a transport matrix (or joint probability) $\boldsymbol{P}$ can be quantified as:

$$d_{\boldsymbol{M}}(\boldsymbol{r},\boldsymbol{c}) = \min_{\boldsymbol{P}\in U(\boldsymbol{r},\boldsymbol{c})} \langle \boldsymbol{P}, \boldsymbol{M}\rangle, \text{ s.t. } \boldsymbol{P}_{1:L} = \boldsymbol{Y}_{1:L} \text{ where}$$
$$U(\boldsymbol{r},\boldsymbol{c}) = \left\{\boldsymbol{P} \in \mathbb{R}_+^{(L+U)\times K} \mid \boldsymbol{P}\mathbf{1} = \boldsymbol{r}, \boldsymbol{P}^\top\mathbf{1} = \boldsymbol{c}\right\}. \tag{11}$$

Algorithm 1: Constrained Wasserstein Mean Shift Clustering (SIAMESE).

Parameters: $\boldsymbol{H}, \boldsymbol{Y}, L, B, N, \lambda, \alpha, n_{\text{step}}, \epsilon$
Initialization: $\tilde{\mathbf{c}}_k = \frac{1}{|S_k|}\sum_{(\boldsymbol{h}_i, y_i)\in S_k} \boldsymbol{h}_i$,
$\boldsymbol{r} = \mathbf{1}, \boldsymbol{c} = (N+B)\mathbf{1}$
while $k < n_{step}$ **do**
  $M_{ij} = \|\boldsymbol{h}_i - \tilde{\boldsymbol{c}}_j\|_2^2, \forall i, j$;
  $\mathbf{P} = \exp(-\lambda \boldsymbol{M})$;
  $\mathbf{P} = \mathbf{P}/\sum_{i,j} P_{ij}$;
  $\boldsymbol{u} = \mathbf{0}_{L+U}$;
  while $\max(|\boldsymbol{u} - \sum_j P_{ij}|) > \epsilon$ **do**
    $\boldsymbol{u} = \sum_j \boldsymbol{P}_{ij}$;
    $\boldsymbol{P} = \text{diag}(\boldsymbol{r}/\boldsymbol{u})\boldsymbol{P}$;
    $\boldsymbol{P} = \boldsymbol{P}\text{diag}(\boldsymbol{c}/\sum_i P_{ij})$;
    $\boldsymbol{P}_{1:L} = \boldsymbol{Y}_{1:L}$;
  end
  $\boldsymbol{\Omega} = \boldsymbol{P}^\top\boldsymbol{H}/(N+B)$;
  $\tilde{\mathbf{C}} \leftarrow \tilde{\mathbf{C}} + \alpha(\boldsymbol{\Omega} - \tilde{\mathbf{C}})$;
  $k \leftarrow k+1$
end
return $y_i = \arg\max_j P_{i,j}$

The optimum of this problem yields distance $d_{\boldsymbol{M}}(\boldsymbol{r},\boldsymbol{c})$. Then we can update $\tilde{\mathbf{C}}$ with $\boldsymbol{P}^\top\boldsymbol{H}/(N+B)$. By alternatively optimizing $\tilde{\mathbf{C}}$ and $\boldsymbol{P}$, we arrive at $\boldsymbol{P}$ which gives us the prediction 'likelihood' for queries. Based on Sinkhorn algorithm [3], we provide our implementation in Alg. 1. Note that $\boldsymbol{Y} \in \mathbb{R}^{(L+U)\times K}$ contains one-hot encoded labels used to preset $\boldsymbol{P}$, that is $\boldsymbol{P}_{1:L} = \boldsymbol{Y}_{1:L}$.

SIAMESE *vs.* Sinkhorn K-means [12]. Sinkhorn K-means [12] is used for the prototype estimation and the so-called partial assignment to prototypes [11]. Note that Sinkhorn K-means is unsupervised *e.g.*, it does not utilize the label information from the support set.
Our minor contribution, SIAMESE, extends Sinkhorn K-means [12] to a semi-supervised setting with label propagation. The line $\boldsymbol{P}_{1:L} = \boldsymbol{Y}_{1:L}$ in Alg. 1 imposes the best possible distribution on labeled support samples. Across all datasets, EASE+SIAMESE was consistently better than EASE+Sinkhorn K-means by 0.3–0.4%.

4. Experiments

We evaluate our approach on four few-shot classification benchmarks, mini-ImageNet [43], tiered-ImageNet [33], CUB [46], CIFAR-FS [2, 18] and OpenMIC [16, 17], used in transductive and semi-supervised FSL works [15, 26, 31, 33, 39]. On these benchmarks, we use the standard evaluation protocols. The results of the transductive and semi-supervised FSL evaluation together with comparison to pre-

Table 1. Test accuracy *vs*. the state of the art (1-shot and 5-shot classification).

			mini-ImageNet		tiered-ImageNet	
Methods	Setting	Network	1-shot	5-shot	1-shot	5-shot
MAML [6]	Inductive	ResNet-18	49.61 ± 0.92	65.72 ± 0.77	–	
RelationNet [41]	Inductive	ResNet-18	52.48 ± 0.86	69.83 ± 0.68	–	
MatchingNet [43]	Inductive	ResNet-18	52.91 ± 0.88	68.88 ± 0.69	–	
ProtoNet [40]	Inductive	ResNet-18	54.16 ± 0.82	73.68 ± 0.65	–	
DeepEMD [50]	Inductive	ResNet-12	65.91 ± 0.82	82.41 ± 0.56		
TPN [26]	transductive	ResNet-12	55.51 ± 0.86	69.86 ± 0.65	59.91 ± 0.94	73.30 ± 0.75
TEAM [31]	transductive	ResNet-18	60.07	75.9	–	
Transductive tuning [4]	Transductive	ResNet-12	62.35 ± 0.66	74.53 ± 0.54	–	
MetaoptNet [21]	Transductive	ResNet-12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
DSN-MR [39]	Transductive	ResNet-12	64.60 ± 0.72	79.51 ± 0.50	67.39 ± 0.82	82.85 ± 0.56
CAN-T [9]	Transductive	ResNet-12	67.19 ± 0.55	80.64 ± 0.35	73.21 ± 0.58	84.93 ± 0.38
EASE+Soft K-means (ours)	Transductive	ResNet-12	57.00 ± 0.26	75.07 ± 0.21	69.74 ± 0.31	85.17 ± 0.21
QR+SIAMESE (ours)	Transductive	ResNet-12	68.66 ± 0.37	79.36 ± 0.22	75.87 ± 0.29	87.80 ± 0.21
EASE+SIAMESE (ours)	Transductive	ResNet-12	70.47 ± 0.30	80.73 ± 0.16	84.54 ± 0.27	89.63 ± 0.15
ProtoNet [40]	Inductive	WRN-28-10	62.60 ± 0.20	79.97 ± 0.14	–	
MatchingNet [43]	Inductive	WRN-28-10	64.03 ± 0.20	76.32 ± 0.16	–	
SimpleShot [44]	Inductive	WRN-28-10	65.87 ± 0.20	82.09 ± 0.14	70.90 ± 0.22	85.76 ± 0.15
S2M2-R [27]	Inductive	WRN-28-10	64.93 ± 0.18	83.18 ± 0.11	-	-
Transductive tuning [4]	Transductive	WRN-28-10	65.73 ± 0.68	78.40 ± 0.52	73.34 ± 0.71	85.50 ± 0.50
SIB [10]	Transductive	WRN-28-10	70.00 ± 0.60	79.20 ± 0.40	–	
BD-CSPN [25]	Transductive	WRN-28-10	70.31 ± 0.93	81.89 ± 0.60	78.74 ± 0.95	86.92 ± 0.63
TIM [1]	Transductive	WRN-28-10	77.8	87.4	82.1	89.8
EPNet [34]	Transductive	WRN-28-10	70.74 ± 0.85	84.34 ± 0.53	78.50 ± 0.91	88.36 ± 0.57
LaplacianShot [58]	Transductive	WRN-28-10	74.86 ± 0.19	84.13 ± 0.14	80.18 ± 0.21	87.56 ± 0.15
iLCT [20]	Transductive	WRN-28-10	83.05 ± 0.79	88.82 ± 0.42	88.50 ± 0.75	92.46 ± 0.42
Oblique Manifold [30]	Transductive	WRN-28-10	80.64 ± 0.34	89.39 ± 0.39	85.22 ± 0.34	91.35 ± 0.40
EASE+Soft K-means (ours)	Transductive	WRN-28-10	67.42 ± 0.27	84.45 ± 0.18	75.87 ± 0.29	85.17 ± 0.21
QR+SIAMESE (ours)	Transductive	WRN-28-10	79.90 ± 0.34	86.88 ± 0.19	84.31 ± 0.30	90.55 ± 0.19
EASE+SIAMESE (ours)	Transductive	WRN-28-10	83.00 ± 0.21	88.92 ± 0.13	88.96 ± 0.23	92.63 ± 0.13

vious methods are summarized in Tables 1, 2, 3, 4, 5 and 6 respectively, and discussed in the following sections. The performance numbers are given as accuracy %, and the 0.95 confidence intervals are reported. The tests are performed on 10,000 random 5-way episodes, with 1 or 5 shots (number of support examples per class), and with 15 queries per episode (2 for OpenMIC). We use publicly available pre-trained backbone convolutional neural networks that are trained on the base class training set. We experiment with the residual network ResNet-12 [29], Wide Residual Network WRN-28-10 [35] and DenseNet [23]. More details about experiments are in the supplementary material.

4.1. FSL benchmarks used in our experiments

Transductive FSL Setting. In this setting, support and query sets provide labeled and unlabeled data. In Table 1, we report the performance of our proposed EASE, SIAMESE and EASE+SIAMESE, and compare them to a set of baselines and state-of-the-art (SOTA) transductive FSL methods from the literature: TPN [26], Transductive FineTuning [4], MetaOptNet [21], DSN-MR [39], EPNet [34], CAN-T [9], SIB [48], BP-CSPN [26], LaplacianShot [58], RAP-LaplacianShot [8], ICI [45], TIM [1], iLPC [20], PT-MAP [11] and ConstellationNet [48]. We also compare to SOTA regular FSL result of S2M2-R [27] in order to highlight the effect of using the unlabeled queries for prediction. Tables 1, 2 and 3 (mini-ImageNet and tiered-ImageNet) show that EASE (best variant) outperforms all the previous (transductive/inductive) SOTA. Our variants include EASE+Soft K-means and QR+SIAMESE. Soft K-means is standard in transductive (and semi-supervised) FSL [39, 40], and QR is a matrix decomposition [42].

Some recent models such as MCT [19] use data/model perturbation/augmentations and achieve results even better than ours. However, we use the TAFSSL protocol (and thus their features) for the common testbed with other methods. On mini-Imagenet (1- and 5-shot protocols), MCT without perturbations (ResNet-12, transductive) yields 71.95% and 81.06% *vs*. our 71.5% and 81.37% (for $n_{step} = 50$, $\epsilon = 0.0001$ in Alg. 1). On tired-Imagenet, MCT (with aug-

Table 2. Test accuracy *vs*. the state of the art based on transductive inference (1-shot and 5-shot classification on CUB).

Method	Backbone	1-shot	5-shot
LaplacianShot [58]	ResNet18	80.96	88.68
LR+ICI [45]	ResNet12	86.53±0.79	92.11±0.35
iLPC [20]	ResNet12	89.00±0.70	92.74±0.35
EASE+Soft K-means (ours)	ResNet-12	76.72±0.27	90.04±0.16
QR+SIAMESE (ours)	ResNet-12	85.60±0.27	91.11±0.16
EASE+SIAMESE (ours)	ResNet12	90.11±0.21	93.13±0.11
BD-CSPN [25]	WRN-28-10	87.45	91.74
TIM-GD [1]	WRN-28-10	88.35±0.19	92.14±0.10
PT+MAP [11]	WRN-28-10	91.37±0.61	93.93±0.32
LR+ICI [45]	WRN-28-10	90.18±0.65	93.35±0.30
iLPC [20]	WRN-28-10	91.03±0.63	94.11±0.30
EASE+Soft K-means (ours)	WRN-28-10	81.01±0.26	91.44±0.14
QR+SIAMESE (ours)	WRN-28-10	89.96±0.26	93.09±0.14
EASE+SIAMESE (ours)	WRN-28-10	91.68±0.19	94.12±0.09

Table 3. Test accuracy *vs*. the state of the art based on transductive inference (1-shot and 5-shot classification on CIFAR-FS).

Method	Backbone	1-shot	5-shot
LR+ICI [45]	ResNet-12	75.36±0.97	84.57±0.57
iLPC [20]	ResNet-12	77.14±0.95	85.23±0.55
DSN-MR [39]	ResNet-12	75.60±0.90	86.20±0.60
ConstellationNet [48]	ResNet-12	75.40±0.20	86.80±0.20
EASE+Soft K-means (ours)	ResNet-12	63.98±0.21	81.05±0.16
QR+SIAMESE (ours)	ResNet-12	75.46±0.26	84.50±0.16
EASE+SIAMESE (ours)	ResNet-12	78.41±0.29	85.67±0.11
SIB [10]	WRN-28-10	80.00±0.60	85.30±0.40
PT+MAP [11]	WRN-28-10	86.91±0.72	90.50±0.49
LR+ICI [45]	WRN-28-10	84.88±0.79	89.75±0.48
iLPC [20]	WRN-28-10	86.51±0.75	90.60±0.48
EASE+Soft K-means (ours)	WRN-28-10	75.61±0.20	87.64±0.15
QR+SIAMESE (ours)	WRN-28-10	86.00±0.22	89.82±0.22
EASE+SIAMESE (ours)	WRN-28-10	87.60±0.23	90.60±0.16

Table 4. Test accuracy *vs*. the state of-the art (1-shot classification, transductive, OpenMIC). All baselines are inductive.

Model	5-way 1-shot				
	p1→p2	p2→p3	p3→p4	p4→p1	Avg
Matching Nets [43]	69.40	57.30	76.35	53.68	64.18
Relation Nets [41]	70.10	49.70	66.90	46.90	58.40
Prototypical Nets [40]	66.33	52.03	74.28	54.30	61.74
SoSN [51]	78.00	60.10	75.50	57.80	67.85
DSN [39]	75.87	62.13	78.25	62.11	69.59
EASE+SIAMESE (ours)	81.60	68.68	86.38	66.53	75.80

mentations/perturbations) yields 82.32% and 87.36%. We get 84.51% and 89.64% (no aug. or perturbations).

Semi-supervised Learning. In this setting, one has an access to an additional set of unlabeled samples (along with each test task), which may contain both the target task category and samples of some other category. Table 5 summarizes the performance of our methods *vs*. SOTA semi-supervised FSL methods. As shown in Table 5, our method is superior to other competitors in all settings by a significant margin (ResNet-12 backbone) *e.g*., the gain varies between 3% and 6% on mini-ImageNet 1-shot protocol. This can be attributed to improved capturing of the data-manifold structure given extra unlabeled samples. With the backbone of WRN-28-10, our method also outperforms other methods in the one-shot setting *e.g*., between 1.3% and 3.5% on mini-ImageNet 1-shot protocol. As PT+MAP [11] does not list semi-supervised results, we used the PT+MAP implementation from iLPC [20] with the backbone of WRN-28-10. In the experiments on tiered-ImageNet, the relatively large number of categories resulted in randomly selected very diverse unlabeled samples which have no positive effect on the support and query sets. Thus results are usually worse than the performance in the transductive setting. We also outperform PTN [14] in mini-ImageNet (84.89±0.74 vs 82.66±0.97 in one-shot setting) and tiered-ImageNet (90.08±0.62 vs 84.70±1.14).

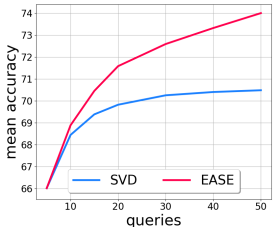

Figure 2. Number of queries in the transductive FSL setting on mini-ImageNet.

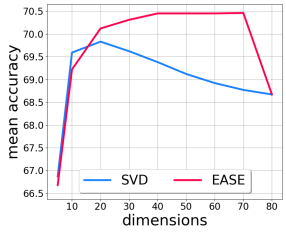

Figure 3. Dimension reduction in the transductive FSL setting on mini-ImageNet.

Our method is insensitive to the feature extractor. To this end, we report the performance of EASE with DenseNet in Table 6, which trains the backbone (from scratch) with a regular multi-class classifier on all training classes rather than in the meta-learning setting. Compared with the baseline SimpleShot, our method gains 13% in the 1-shot setting and more than 4% in the 5-shot setting. The proposed method also outperforms other transductive methods based on DenseNet such as LaplacianShot [58], RAP-LaplacianShot [8] and TAFSSL [23].

4.2. Ablations

Number of queries in transductive FSL. Since the unlabeled data in transductive FSL is comprised entirely from the query samples, the size of the query set in the meta-testing episodes affects the performance. To test its effect, we evaluate the proposed variants of EASE. All methods were tested by varying the number of queries from 2 to 50. Figure 2 (mini-ImageNet) shows that for 5 queries or more, our method is effective in learning a discriminative subspace from unlabeled samples. As the number of queries increases, the performance of the SVD-based transductive inference soon stops improving. However, our method still gains on the performance approximately linearly.

Table 5. Comparison of test accuracy against state-of-the-art methods for 1-shot and 5-shot classification under the semi-supervised few-shot learning setting. CUB 5-shot omitted: no class has the required 70 examples.

			mini-ImageNet		tiered-ImageNet		CIFAR-FS		CUB	
Methods	Network	Setting	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
LR+ICI [45]	ResNet-12	30/50	$67.57_{\pm 0.97}$	$79.07_{\pm 0.56}$	$83.32_{\pm 0.87}$	$89.06_{\pm 0.51}$	$75.99_{\pm 0.98}$	$84.01_{\pm 0.62}$	$88.50_{\pm 0.71}$	-
iLPC [20]	ResNet-12	30/50	$70.99_{\pm 0.91}$	$81.06_{\pm 0.49}$	$85.04_{\pm 0.79}$	$89.63_{\pm 0.47}$	$78.57_{\pm 0.80}$	$85.84_{\pm 0.56}$	$90.11_{\pm 0.64}$	-
EASE+SIAMESE (ours)	ResNet-12	30/50	$73.90_{\pm 0.29}$	$81.68_{\pm 0.48}$	$85.86_{\pm 0.79}$	$89.64_{\pm 0.45}$	$80.51_{\pm 0.88}$	$85.96_{\pm 0.54}$	$90.61_{\pm 0.61}$	-
LR+ICI [45]	WRN-28-10	30/50	$81.31_{\pm 0.84}$	$88.53_{\pm 0.43}$	$88.48_{\pm 0.67}$	$92.03_{\pm 0.43}$	$86.03_{\pm 0.77}$	$89.57_{\pm 0.53}$	$90.82_{\pm 0.59}$	-
PT+MAP [11]	WRN-28-10	30/50	$83.14_{\pm 0.72}$	$88.95_{\pm 0.38}$	$89.16_{\pm 0.61}$	$92.30_{\pm 0.39}$	$87.05_{\pm 0.69}$	$89.98_{\pm 0.49}$	$91.52_{\pm 0.53}$	-
iLPC [20]	WRN-28-10	30/50	$83.58_{\pm 0.79}$	$89.68_{\pm 0.37}$	$89.35_{\pm 0.68}$	$92.61_{\pm 0.39}$	$87.03_{\pm 0.72}$	$90.34_{\pm 0.50}$	$91.69_{\pm 0.55}$	-
EASE+SIAMESE (ours)	WRN-28-10	30/50	$84.89_{\pm 0.74}$	$89.47_{\pm 0.37}$	$90.08_{\pm 0.62}$	$92.67_{\pm 0.39}$	$87.89_{\pm 0.70}$	$90.18_{\pm 0.50}$	$92.11_{\pm 0.52}$	-

Table 6. Test accuracy *vs*. the state of the art (DenseNet, 1-shot and 5-shot classification). The asterisk '*' indicates inductive setting.

	mini-ImageNet		tiered-ImageNet	
Methods	1-shot	5-shot	1-shot	5-shot
SimpleShot* [44]	65.77 ± 0.19	82.23 ± 0.13	71.20 ± 0.22	86.33 ± 0.15
LaplacianShot [58]	75.57 ± 0.19	84.72 ± 0.13	80.30 ± 0.20	87.93 ± 0.15
RAP-LaplacianShot [8]	75.58 ± 0.20	85.63 ± 0.13	-	-
TAFSSL(PCA) [23]	70.53 ± 0.25	80.71 ± 0.16	80.07 ± 0.25	86.42 ± 0.17
TAFSSL(ICA) [23]	72.10 ± 0.25	81.85 ± 0.16	80.82 ± 0.25	86.97 ± 0.17
EASE+Soft K-means	74.30 ± 0.26	82.08 ± 0.17	82.67 ± 0.25	87.60 ± 0.17
QR+SIAMESE	75.75 ± 0.32	85.10 ± 0.18	82.87 ± 0.33	89.11 ± 0.20
EASE+SIAMESE (ours)	79.42 ± 0.27	86.76 ± 0.14	86.17 ± 0.25	90.54 ± 0.15

Compared with SVD (TAFFSL (ICA) without the mean subtraction), EASE shows consistent improvements as we increase the number of queries. The comparison is based on the setting of [23] and uses DenseNet [13]. Note we only use the support set to estimate the class mean for the NCM classifier to avoid interference from the queries.

EASE *vs*. other dimensionality reduction methods. TAFSSL [23] shows that the dimensionality reduction is essential in the transductive inference. Our method can be considered as a dimensionality reduction method because we learn a subspace that improves feature discrimination. To evaluate EASE, we place it at where PCA and ICA of TASSFL are wired. Table 6 shows that our method is superior to counterparts with PCA and ICA in all settings, outperforming them by a significant margin (DenseNet backbone) *e.g*., 3.7% and 2.2% on mini-ImageNet 1-shot protocol. We also find that the larger output dimension of SVD does not always improve the performance. Figure 3 shows that results gradually improve until they peak at 20 (dimension) and then continue to decrease, whereas our method helps improve the performance until the number of dimensions equals the number of samples.

Subspace dimension of EASE. Our method is not very sensitive to the dimension size. Figure 3 shows that its performance improves rapidly between dimension size 5 and 20, and then improves slowly in a relatively stable range. In order to balance the computational effort and performance, 40 dimensions are used throughout experiments.

Inference time. The computational complexity of EASE depends only on the feature dimension and the number of samples. Table 7 provides the average inference time for the 1-shot and 5-shot tasks in mini-ImageNet with the backbones of ResNet-12 and WRN-28-10. On average, our inference takes only a few of milliseconds, which does not impose any burden on the regular usage scenario. Our method is about 10× faster (AMD 2700 CPU) than other two SOTA methods.

Table 7. Average inference time (in seconds) for the 1-shot and 5-shot tasks in mini-ImageNet dataset with different backbones.

BackBone	ResNet12		WRN-28-10	
Shot	1	5	1	5
EASE+SIAMESE	6.0e-3	7.4e-3	8.0e-3	8.7e-3
iLPC [20]	4.5e-2	5.6e-2	5.5e-2	7.0e-2
ICI [45]	3.4e-2	4.2e-2	4.1e-2	5.2e-2

5. Conclusions

Without meta-learning, we provide state-of-the-art results, outperforming significantly a large number of sophisticated few-shot learning methods. The proposed methods are plug-and-play modules for the inference step of few-shot learning as our transductive inference fits into the code of the standard prototypical network. Our solution is simple, efficient and also compatible with semi-supervised approaches. It consists of two components: unsupErvised discriminAnt Subspace lEarning (EASE) and a minor contribution, conStraIned wAsserstein MEan Shift clustEring (SIAMESE). Both components can operate independently. EASE learns a discriminant subspace to minimize the surrogate problem exploiting the data structure without any label information. SIAMESE uses the labeled support set with labels and unlabeled queries to estimate the class means more effectively and improve the final prediction.

Acknowledgements. Hao Zhu is supported by an Australian Government Research Training Program (RTP) Scholarship. Piotr Koniusz is supported by CSIRO's Machine Learning and Artificial Intelligence Future Science Platform (MLAI FSP).

References

[1] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, Jose Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33, 2020. 6, 7

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 2, 5

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013. 5

[4] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019. 1, 2, 6

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 1

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 6

[7] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21–30, 2019. 2

[8] Jie Hong, Pengfei Fang, Weihao Li, Tong Zhang, Christian Simon, Mehrtash Harandi, and Lars Petersson. Reinforced attention for few-shot learning and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 913–923, 2021. 6, 7, 8

[9] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *arXiv preprint arXiv:1910.07677*, 2019. 1, 6

[10] Shell Xu Hu, Pablo G Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil D Lawrence, and Andreas Damianou. Empirical bayes transductive meta-learning with synthetic gradients. *arXiv preprint arXiv:2004.12696*, 2020. 1, 6, 7

[11] Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. *arXiv preprint arXiv:2006.03806*, 2020. 3, 5, 6, 7, 8

[12] Gabriel Huang, Hugo Larochelle, and Simon Lacoste-Julien. Are few-shot learning benchmarks too simple? solving them without task supervision at test-time. *arXiv preprint arXiv:1902.08605*, 2019. 3, 5

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 8

[14] Huaxi Huang, Junjie Zhang, Jian Zhang, Qiang Wu, and Chang Xu. Ptn: A poisson transfer network for semi-supervised few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1602–1609, 2021. 7

[15] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2019. 1, 3, 5

[16] Piotr Koniusz, Yusuf Tas, Hongguang Zhang, Mehrtash Harandi, Fatih Porikli, and Rui Zhang. Museum exhibit identification challenge for the supervised domain adaptation and beyond. In *The European Conference on Computer Vision (ECCV)*, September 2018. 5

[17] Piotr Koniusz and Hongguang Zhang. Power normalizations in fine-grained image, few-shot image and graph classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 5

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[19] Seong Min Kye, Hae Beom Lee, Hoirin Kim, and Sung Ju Hwang. Meta-learned confidence for few-shot learning. *arXiv preprint arXiv:2002.12017*, 2020. 6

[20] Michalis Lazarou, Tania Stathaki, and Yannis Avrithis. Iterative label cleaning for transductive and semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8751–8760, 2021. 5, 6, 7, 8

[21] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019. 6

[22] Xinzhe Li, Qianru Sun, Yaoyao Liu, Shibao Zheng, Qin Zhou, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification (6 2019). In *Advances in Neural Information Processing Systems: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, December*, volume 8, pages 1–11, 1906. 2, 3

[23] Moshe Lichtenstein, Prasanna Sattigeri, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Tafssl: Task-adaptive feature sub-space learning for few-shot classification. In *European Conference on Computer Vision*, pages 522–539. Springer, 2020. 2, 5, 6, 7, 8

[24] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 663–670, 2010. 4

[25] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. *arXiv preprint arXiv:1911.10713*, 2019. 1, 3, 6, 7

[26] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018. 1, 2, 3, 5, 6

[27] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2218–2227, 2020. 6

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and

phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 5

[29] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018. 6

[30] Guodong Qi, Huimin Yu, Zhaohui Lu, and Shuzhao Li. Transductive few-shot classification on the oblique manifold. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8412–8422, 2021. 6

[31] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3603–3612, 2019. 1, 5, 6

[32] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2017. 2

[33] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 2, 3, 5

[34] Pau Rodríguez, Issam Laradji, Alexandre Drouin, and Alexandre Lacoste. Embedding propagation: Smoother manifold for few-shot classification. In *European Conference on Computer Vision*, pages 121–138. Springer, 2020. 2, 6

[35] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018. 6

[36] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 3

[37] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018. 2, 3

[38] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. Meta-learning for multi-label few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3951–3960, 2022. 1

[39] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4136–4145, 2020. 2, 3, 5, 6, 7

[40] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 1, 2, 3, 6, 7

[41] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 2, 6, 7

[42] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997. 6

[43] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016. 2, 5, 6, 7

[44] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 2, 6, 8

[45] Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12836–12845, 2020. 6, 7, 8

[46] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 5

[47] Yaochen Xie, Zhao Xu, Zhengyang Wang, and Shuiwang Ji. Self-supervised learning of graph neural networks: A unified review. *CoRR*, abs/2102.10757, 2021. 5

[48] Weijian Xu, yifan xu, Huaijin Wang, and Zhuowen Tu. Attentional constellation nets for few-shot learning. In *International Conference on Learning Representations*, 2021. 6, 7

[49] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020. 2

[50] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020. 6

[51] Hongguang Zhang and Piotr Koniusz. Power normalizing second-order similarity network for few-shot learning. In *Winter Conference on Applications of Computer Vision (WACV)*, January 2019. 7

[52] Hongguang Zhang, Piotr Koniusz, Songlei Jian, Hongdong Li, and Philip H. S. Torr. Rethinking class relations: Absolute-relative supervised and unsupervised few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9432–9441, 2021. 1

[53] Hongguang Zhang, Hongdong Li, and Piotr Koniusz. Multi-level second-order few-shot learning. *IEEE Transactions on Multimedia*, 2022. 1

[54] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2770–2779, 2019. 1

[55] Shan Zhang, Dawei Luo, Lei Wang, and Piotr Koniusz. Few-shot object detection by second-order pooling. In *Asian Conference on Computer Vision*, 2020. 1

[56] Hao Zhu and Piotr Koniusz. Refine: Random range finder for network embedding. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3682–3686, 2021. 5

[57] Hao Zhu, Ke Sun, and Piotr Koniusz. Contrastive laplacian eigenmaps. *Advances in Neural Information Processing Systems*, 34, 2021. 5

[58] Imtiaz Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *International Conference on Machine Learning*, pages 11660–11670. PMLR, 2020. 5, 6, 7, 8