Semi-Supervised Few-Shot Learning from A Dependency-Discriminant Perspective

Zejiang Hou, Sun-Yuan Kung
Princeton University
{zejiangh,kung}@princeton.edu

Abstract

We study the few-shot learning (FSL) problem, where a model learns to recognize new objects with extremely few labeled training data per category. Most of previous FSL approaches resort to the meta-learning paradigm, where the model accumulates inductive bias through learning from many training tasks, in order to solve new unseen few-shot tasks. In contrast, we propose a simple semi-supervised FSL approach to exploit unlabeled data accompanying the few-shot task to improve FSL performance. More exactly, to train a classifier, we propose a Dependency Maximization loss based on the Hilbert-Schmidt norm of the cross-covariance operator, which maximizes the statistical dependency between the embedded feature of the unlabeled data and their label predictions, together with the supervised loss over the support set. The obtained classifier is used to infer the pseudo-labels of the unlabeled data. Furthermore, we propose an Instance Discriminant Analysis to evaluate the credibility of the pseudo-labeled examples and select the faithful ones into an augmented support set, which is used to retrain the classifier. We iterate the process until the pseudo-labels of the unlabeled data becomes stable. Through extensive experiments on four widely used few-shot classification benchmarks, including mini-ImageNet, tiered-ImageNet, CUB, and CIFARFS, the proposed method outperforms previous state-of-the-art FSL methods.

1. Introduction

Deep learning has achieved remarkable performance on visual recognition problems such as image classification. However, the success of deep neural networks hinges on substantial labeled training examples. The prohibitive annotation cost on very large-scale supervised dataset will limit the applicability of these deep learning models to learn new concepts quickly and efficiently. In contrast, human intelligence has the ability to learn new concepts quickly, even with very few labeled examples. This is achieved by using the prior experience and integrating it with the new information. By the same token, it is desirable for the deep learning models to learn to recognize novel classes of objects with very limited labeled examples. This learning approach is referred to as the few-shot learning (FSL), which has received substantial research interests.

A large body of FSL approaches place focus on the meta-learning paradigm and episodic training strategy. In meta-learning, the model is trained on a series of episodes, with support and query examples, that simulate the generalization during testing time. After accumulating the prior experience, the trained model will have the ability to generalize to novel classes by using very few labeled training data. However, a recent work [5] empirically found that meta-learning cannot compete with the simplest transfer learning baseline, where a deep backbone model is firstly trained on a big supervised image corpus and a new linear classifier is appended and finetuned on the novel few-shot tasks.

Recent methods start exploring transductive and semi-supervised learning to improve the performance on few-shot tasks, by using the information from unlabeled query examples or an additional unlabeled set. Among various methods, self-training [19] is one of the most straightforward way to utilize the unlabeled data. Typically, a model trained on the support examples can be used to infer the pseudo-labels (class that has the maximum predicted probability) of the unlabeled data, and then uses these pseudo-labels along with the support set to retrain the model. However, in FSL, the model is trained with very few labeled support examples, thus it cannot capture the data distribution of target classes in the task. The inferred. pseudo-labels will have low quality. Including wrongly labeled examples into the training set will jeopardize the final model performance.

To counteract the limitations in meta-learning and traditional self-training, we present a novel semi-supervised FSL approach to exploit the unlabeled examples to improve few-shot performance. Firstly, we propose a *Dependency Maximization* loss to enhance the model training, which maximizes the statistical dependence between the embed-

ded features of unlabeled data and their softmax predictions, in conjunction with the supervised loss minimization over support set. We develop a empirical dependence measure based on the Hilbert-Schmidt norm of the cross-covariance operator. The obtained model is used to infer the pseudo-labels for those unlabeled data, where we further propose an *Instance Discriminant Analysis* to evaluate the sample from the perspective of feature discriminant power and select the most faithful pseudo-labels to augment the support set and retrain the model. Following the standard transductive and semi-supervised FSL setup, our experiments show that the proposed method outperforms previous state-of-the-art FSL methods, not only on the widely adopted few-shot benchmarks, but on more challenging scenarios such as cross-domain FSL and higher-way testing FSL.

Our contribution are highlighted as follows:

- A simple semi-supervised few-shot learning framework to exploit unlabeled data in few-shot tasks
- A dependency maximization loss to train the classifier with unlabeled data
- An instance discriminant analysis method to evaluate and select credible pseudo-labels. We also derive an efficient approximation for our discriminant criterion to speed up our selection process substantially.
- State-of-the-art performance on various few-shot classification benchmarks

2. Related Works

We briefly review recently proposed few-shot learning approaches, with focus on transductive and semi-supervised FSL methods. Optimization-based meta-learning methods [1,7,23,26] learn the model through a series of episodes, so that it can adapt to new tasks with limited labeled examples. In contrast, our method does not use the complex meta-training setup; we pretrain a feature extractor on the base classes using the standard cross-entropy loss. Metric learning methods learn to compare feature similarity based on distance metric between support and query examples in the feature space. Examples of the distance metrics include cosine similarity [29], Euclidean distance [25, 32], relation network [10, 27], mahalanobis distance [3], Earth Mover's distance [34], subspace projection distance [24]. In this paper, we do not utilize any specialized distance metric, instead we propose a label-free dependency maximization loss for task inference. Hallucination based methods [8, 13, 35] utilize generative models or data augmentations to expand the support set by synthesizing new samples or features based on the given labeled data. In contrast, our framework uses the additional unlabeled data using pseudo-labeling technique.

Transductive and Semi-Supervised FSL. In practical applications, we can access to unlabeled data accompanying the few-shot task, apart from the labeled support set. Transductive FSL (TFSL) methods [4, 6, 11, 17, 22] assume that the query examples come in as a bulk and can be used as unlabeled data to facilitate the few-shot performance. [17] uses label-propagation to propagate labels from labeled to unlabeled examples via a graph. [22] proposes embedding-propagation regularizer for manifold smoothing. [11] proposes a Laplacian regularizer to encourage nearby query samples to have consistent label assignments. [6] proposes to minimize the conditional entropy of the query softmax predictions. Similarly, [4] further incorporates a marginal entropy of the query softmax predictions, which helps to avoid degenerate solutions obtained when solely minimizing conditional entropy. In contrast, our method proposes to maximize the statistical dependency between the features and their label predictions. In semi-supervised FSL (SSFSL), the unlabeled data comes in addition to the support/query set. To name a few, [14] applies self-labeling and soft-attention to the unlabeled set with finetuning on both labeled and self-labeled examples. [21] proposes a prototype refinement based on the soft assignment scores for the unlabeled examples. [31] introduces a linear regression hypothesis to select pseudo-labeled examples for classifier training. Different from these approaches, we propose a simple instance discriminant analysis, together with our dependency maximization loss, to utilize the unlabeled data for improving FSL performance.

3. Preliminary

3.1. Few-shot learning formulation

Assume we are given a labeled base dataset $\mathcal{X}_{base} = \{(\mathbf{x}_i, \mathbf{y}_i), \mathbf{y}_i \in \mathcal{Y}_{base}\}$, where $\mathcal{Y}_{base}$ denotes the set of classes (i.e. category set) for the base dataset. FSL aims to learn a model using the base dataset so that the model will be able to recognize **unseen** examples belonging to **novel classes** after short learning from very few labeled examples in these novel classes. Denote a novel dataset by $\mathcal{X}_{novel} = \{(\mathbf{x}_i, \mathbf{y}_i), \mathbf{y}_i \in \mathcal{Y}_{novel}\}$ with completely new category set $\mathcal{Y}_{novel}$, from which the few-shot tasks are sampled. The base and novel datasets have mutually disjoint classes, i.e. $\mathcal{Y}_{base} \cap \mathcal{Y}_{novel} = \emptyset$. We follow the standard N-way K-shot formulation. For each few-shot task $\mathcal{T}_i$, we randomly sample N classes from $\mathcal{Y}_{novel}$. We then sample K labeled examples for each of N classes and construct the *support set* $\mathcal{D}^S_{\mathcal{T}_i}$ with set size $|\mathcal{D}^S_{\mathcal{T}_i}| = N \times K$. Each task also has a *query set* $\mathcal{D}^Q_{\mathcal{T}_i}$, which consists of Q unlabeled and unseen examples for the same N classes, i.e. $|\mathcal{D}^Q_{\mathcal{T}_i}| = Q \times K$. The unlabeled query set evaluates the generalization performance of the model trained on base set and also adapted on the labeled support set.

3.2. Self-training based semi-supervised FSL

One of the fundamental challenge for FSL is the difficulty to estimate the data distribution of novel categories with very few labeled examples. To address this problem, recent FSL approaches use semi-supervised learning (SSFSL) or transductive learning (TFSL) wit h unlabeled examples $\mathcal{D}^U_{\mathcal{T}_i}$ of novel categories for the task at hand. In SSFSL setting, extra examples apart from the support and query examples are available for the model to learn, while TFSL assumes the model evaluates all query examples at once and utilizes those query examples as the unlabeled set. Among various semi-supervised learning methods, self-training [19] is one of the state-of-the-art representatives that can be easily applied. Specifically, let $f_\theta : \mathcal{X} \to \mathcal{Z} \subset R^d$ be the feature extractor of a deep neural network parameterized by θ, where $\mathcal{Z}$ denotes the space of the feature embedding. In this work, we pretrain the feature-extractor on the base dataset $\mathcal{X}_{base}$, following [4]. Given a few-shot task $\mathcal{T}_i$, self-training first learns a classifier from the labeled support set: $\min_{\phi} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}^S_{\mathcal{T}_i}} \mathcal{L}(h_\phi(f_\theta(\mathbf{x})), \mathbf{y})$ where $\mathcal{L}$ represents the cross-entropy loss. Then, the classifier is used to infer the pseudo-labels $\hat{y}_u = h_\phi(f_\theta(\mathbf{x}_u))$ for the unlabeled examples $\mathcal{D}^U_{\mathcal{T}_i} = \{\mathbf{x}_u\}^U_{u=1}$. The pseudo-labeled examples are taken as additional labeled data for the corresponding classes and are augmented with the support examples using their pseudo-labels as true labels. Finally, the classifier is retrained using the augmented support set and evaluated on the query set.

3.3. Motivation for our method

Self-training in FSL suffers from two limitations: (1) since the classifier $h_\phi(\cdot)$ is trained with very few labeled examples, the pseudo-labels have low quality with significant label noise; (2) there is no sample selection strategy to filter out the label noise, and including the untrustworthy pseudo-labeled examples for the target classes causes accuracy degradation, counteracting our goal of using unlabeled data to improve the accuracy. To rectify these two limitations, (1) we propose a *Dependency Maximization* loss which gracefully uses the unlabeled data to enhance the classifier training process for generating higher quality pseudo labels; (2) we propose an *Instance Discriminant Analysis* to evaluate the pseudo-labeled examples and select the most trustworthy ones to augment the support set, in order to alleviate the accumulation of label noise.

4. Methodology

4.1. Overview

Given a pretrained feature extractor CNN $f_\theta(\cdot)$, for a specific few-shot learning task with support set $D^S_{\mathcal{T}_i}$ and unlabeled set $D^U_{\mathcal{T}_i}$, the proposed method performs the following steps:

1. Apply an self-attention based feature pre-processing to features in the unlabeled set and support set
2. Train a classifier using the supervised loss on the support set and the dependency maximization loss on the unlabeled set
3. Infer the pseudo-labels on the unlabeled set
4. Evaluate the credibility of the pseudo-labeled examples based on the instance discriminant analysis and select the most trustworthy ones to augment the support set
5. Repeat step 2 - 4 until the pseudo-labels on the unlabeled set becomes stable
6. Use the final classifier to predict the query set $D^Q_{\mathcal{T}_i}$

4.2. Feature pre-processing via self-attention

The feature pre-precessing step is applied to transform both the unlabeled set and the support set. For the unlabeled set, the pre-precessing step captures the global relationships among different instances, so that similar features can be driven closer to induce better data separation. We transform each unlabeled instance by fusing its feature with a weighted sum of other instances' features using self-attention [28]. Let $\mathbf{X}_U = [f_\theta(\mathbf{x}_1), ..., f_\theta(\mathbf{x}_{|D^U_i|})]$ denote the data matrix with the unlabeled features arranged in rows. Then the transformed unlabeled set is given by:

$$\mathbf{X}_U \leftarrow (1-\alpha) * \mathbf{X}_U + \alpha * \text{softmax}\Big(\frac{d(\mathbf{X}_U, \mathbf{X}_U)}{\tau}\Big)\mathbf{X}_U \quad (1)$$

In Eq.(1), α is a constant balancing the transformed and original features. $d(\cdot,\cdot)$ is a distance metric and we compute the squared Euclidean distance between each pair of unlabeled instances in the feature embedding space: $d(\mathbf{X}_U, \mathbf{X}_U) = -(2\text{diag}(\mathbf{X}_U\mathbf{X}_U^T) - 2\mathbf{X}_U\mathbf{X}_U^T)$. The softmax operation is applied row-wise so that the attention weights summed up to one. τ is a temperature constant that controls the sharpness of the attention weights. As the attention weights are normalized similarities between one unlabeled instance's features with all other instances' features, the transformation drives similar features closer to each other.

For the support set, the pre-processing step propagates the information from the unlabeled set to support instance via cross-attention. Let $\mathbf{X}_S = [f_\theta(\mathbf{x}_1), ..., f_\theta(\mathbf{x}_{|D^S_{\mathcal{T}_i}|})]$ denote the data matrix with the labeled support features arranged in rows. Then, the transformed support set is given by:

$$\mathbf{X}_S \leftarrow (1-\beta) * \mathbf{X}_S + \beta * \text{softmax}\Big(\frac{d(\mathbf{X}_S, \mathbf{X}_U)}{\tau}\Big)\mathbf{X}_U \quad (2)$$

In Eq.(2), the distance metric is calculated between the features in the support set and the transformed features in the unlabeled set.

4.3. Dependency maximization for training classifier

The proposed *Dependency Maximization* (DM) loss is differentiable and can be optimized with standard gradient descent algorithm to enhance the classifier training. While we train the classifier with the labeled support examples, DM loss maximizes the statistical dependence between the features of the unlabeled set and their label predictions, in conjunction with minimizing the cross-entropy loss over the support set. DM loss acts as a surrogate for the classifier's empirical risk on the unlabeled examples, which helps to restrict the classifier's hypothesis space and facilitates the prediction for the given unlabeled examples.

We begin by listing some notations before introducing how to characterize the dependence between features and label predictions. Let Z be a random vector (of size $\mathbb{R}^d$ where d is the embedding dimension) associated with the embedded features of the unlabeled set, Y denotes the random vector (of size $\mathbb{R}^N$ where N is number of classes) associated with their softmax predictions. $P_{Z,Y}$ is the joint distribution between these two random variables. To measure the dependence between Z and Y, we define the cross-covariance operator based on [2]:

$$C_{zy} := \mathbb{E}_{zy}[(\Phi(z) - \mu_z) \otimes (\Psi(y) - \mu_y)] \tag{3}$$

where $\Phi : \mathcal{Z} \to \mathcal{F}$ (or $\Psi : \mathcal{Y} \to \mathcal{G}$) defines a kernel mapping from the space of feature embedding (or space of prediction vector) to a reproducing kernel Hilbert space (RKHS) F (or G), with mean vectors defined as μ_z (or μ_y). To summarize the degree of dependence between Z and Y, we use the Hilbert-Schmidt norm of C_{zy}, which is given by the trace of $C_{zy}C_{zy}^T$. In this paper, we use the square of the Hilbert-Schmidt norm of the cross-covariance operator, $\|C_{zy}\|_{HS}^2$, because it can detect nonlinear dependence, as shown in the following theorem:

Theorem 1 (C_{zy} and Independence [9]). *Assume F and G are RKHSs with characteristic kernels. Then, we have $\|C_{zy}\|_{HS}^2 = 0$ if and only if Z and Y are independent.*

Characteristic kernels such as Gaussian kernel, i.e. $k(x, x') = \exp\big(-\|x - x'\|_2^2/(2\sigma^2)\big)$, allows us to measure any dependence between Z and Y. In our case, $\|C_{zy}\|_{HS}^2$ is zero only if the features and the label predictions of the unlabeled set are independent. Clearly, we aim to achieve the opposite. We aim to maximize the dependence between the features and predictions by maximizing $\|C_{zy}\|_{HS}^2$.

To utilize the dependence measure as a loss function for classifier training, we need an empirical estimate from finite number of samples. Formally, denote the kernel functions associated with the RKHS F and G as $k(z, z')$ and $l(y, y')$; let $\mathbf{K}, \mathbf{L} \in \mathbb{R}^{U\times U}$ denote the Gram matrices defined over the features and softmax predictions associated with the unlabeled set $\mathcal{D}_{\mathcal{T}_i}^U$, containing entries $\mathbf{K}_{i,j} = k(z_i, z_j)$ and $\mathbf{L}_{i,j} = l(y_i, y_j)$. The empirical estimator of $\|C_{zy}\|_{HS}^2$ is given by:

$$\widehat{\|C_{zy}\|_{HS}^2} := (U-1)^{-2}\,\mathrm{tr}(\mathbf{KHLH}) \tag{4}$$

where $\mathbf{H} = \mathbf{I}_U - (1/U)\mathbf{1}_U\mathbf{1}_U^T$ is the centering matrix, $\mathbf{I}_U$ is an identity matrix, $\mathbf{1}_U$ is a vector with all ones, and $\mathrm{tr}(\cdot)$ is the matrix trace operation. As shown by [9], this empirical estimator converges sufficiently: with high probability, $|\widehat{\|C_{zy}\|_{HS}^2} - \|C_{zy}\|_{HS}^2|$ is bounded by a very small constant.

With the empirical estimator, we define the overall loss function, by adding the empirical dependence measure on the unlabeled set with the supervised cross-entropy loss on the support set:

$$\min_{\mathbf{W},\mathbf{b}} \underbrace{-\frac{1}{NK}\sum_{(x,y)\in\mathcal{D}_{\mathcal{T}_i}^S}\log\frac{\exp(\mathbf{W}_y^T f_\theta(\mathbf{x}) + \mathbf{b}_y)}{\exp(\sum_{c=1}^N \mathbf{W}_c^T f_\theta(\mathbf{x}) + \mathbf{b}_c)}}_{\text{Cross-entropy minimization on support set}} \underbrace{-\lambda\cdot(U-1)^{-2}\,\mathrm{tr}(\mathbf{KHLH})}_{\text{Dependency maximization on unlabeled set}} \tag{5}$$

where $\mathbf{W}, \mathbf{b}$ denote the weight and bias of the softmax linear classifier h_ϕ, and the label prediction is given by $\hat{\mathbf{y}} = h_\phi(\mathbf{z}) = \mathrm{softmax}(\mathbf{W}^T\mathbf{z} + \mathbf{b}) = \mathrm{softmax}(\mathbf{W}^T f_\theta(\mathbf{x}) + \mathbf{b})$.

Objective (5) is optimized for each test task using gradient descent (GD) w.r.t. $\mathbf{W}$ and $\mathbf{b}$. The pretrained feature extractor f_θ is frozen. $\mathbf{W}$ and $\mathbf{b}$ are initialized based on the class prototypes computed over the support set: $\mathbf{W}^0 = [2\boldsymbol{\mu}_1, ..., 2\boldsymbol{\mu}_N] \in \mathbb{R}^{d\times N}$ and $\mathbf{b}^0 = [-\|\boldsymbol{\mu}_1\|_2^2, ..., -\|\boldsymbol{\mu}_N\|_2^2] \in \mathbb{R}^{N\times 1}$. Then, the weight and bias parameters are updated by GD using both the support and unlabeled samples of the few-shot task without mini-batch sampling.

4.4. Instance discriminant analysis for selecting pseudo-label

After training the classifier with our proposed DM loss in Eq.(5), we can predict the labels $\hat{y}_u$ for the unlabeled examples in $\mathcal{D}_{\mathcal{T}_i}^U$ as their pseudo-labels. In this section, we present an *Instance Discriminant Analysis* (IDA) to evaluate the quality of these pseudo-labeled examples and select the most trustworthy ones into the augmented support set.

IDA is essentially a sample selection or outlier removal algorithm, that aims to remove a subset of training data sample *a priori*, and train the classifier only with the remaining subset of data. We introduce a hypothesis that the feature discriminant power computed on the embedded features and pseudo-labels can be used as a surrogate for the pseudo-labels' accuracy for the unlabeled set (cf. Table 1). The rationale behind is that a wrongly labeled example would be detrimental to the overall data separability of the unlabeled set, causing low feature discriminant power; while a correctly labeled example would facilitate the data separability, improving the feature discriminant power.

We adopt Fisher's discriminant analysis as the basis of our quality measure. We evaluate the quality of each

Labeling	*mini*-ImageNet		*tiered*-ImageNet	
	ψ	Acc.(%)	ψ	Acc.(%)
Random	0.14	20.00	0.13	20.00
Pseudo (w/o DM)	0.55	57.73	0.61	68.29
Pseudo (w/ DM)	0.68	75.80	0.74	82.43
Groundtruth	1.00	100.00	1.00	100.00

Table 1. The feature discriminant (measured by the normalized ψ in Eq.(6)) can serve as a surrogate for the pseudo-labels' actual accuracy over the unlabeled set. Thus, ψ in Eq.(6) can be used as a metric to evaluate the credibility of the pseudo-labels. "Random": randomly guessing the labels for unlabeled data. "w/o DM": inferred pseudo-labels from a classifier trained with cross-entropy loss only. "w/ DM": inferred pseudo-labels from a classifier trained with cross-entropy and dependency loss as in Eq.(5).

Algorithm 1 Semi-supervised FSL from A Dependency-Discriminant Perspective

Require Support set $D^S_{\mathcal{T}_i} = \{\mathbf{x}_n, y_n\}_{n=1}^{NK}$; Unlabeled data $D^U_{\mathcal{T}_i} = \{\mathbf{x}_u\}_{u=1}^{U}$; pretrained feature extractor $f_\theta(\cdot)$.
Run self-attention feature preprocessing on the support and unlabeled set as Eq.(1) and (2).
Initialize augmented support set $(X_s, y_s) = \{\mathbf{x}_n, y_n\}_{n=1}^{NK}$.
while pseudo-labels are not stablized **do**
    Train a classifier $h_\phi(\cdot)$ on (X_s, y_s) and $D^U_{\mathcal{T}_i}$ using Eq.(5).
    Infer pseudo-labels for $\{\mathbf{x}_u\}_{u=1}^{U}$ and obtain $\{\mathbf{x}_u, \hat{y}_u\}_{u=1}^{U}$.
    Compute IDA for each pseudo-labeled instance by Eq.(7).
    Rank $\{\mathbf{x}_u, \hat{y}_u\}_{u=1}^{U}$ based on their IDA values $d\psi_u$.
    Select the most trustworthy subset (X_{sub}, y_{sub}) from $\{\mathbf{x}_u, \hat{y}_u\}_{u=1}^{U}$, and merge them into (X_s, y_s).
end while
Return Augmented support set (X_s, y_s).

pseudo-labeled instance by computing its contribution to the overall data separability based on the *Fishers Criterion*. Formally, denote the set of pseudo-labeled data as $\{(\mathbf{x}_u, \hat{y}_u)|\mathbf{x}_u \in \mathcal{D}^U_{\mathcal{T}_i}\}$. Note that the true labels are not known, but we use the pseudo labels for the following computation. Let $f(\mathbf{x}_u)$ denote the embedded feature of instance u (for notation simplicity, we omit θ for the feature extractor). We define the *scatter matrix* and the *between-class scatter matrix*, respectively, as $\bar{\mathbf{S}} = \sum_{u=1}^{U}(f(\mathbf{x}_u) - \boldsymbol{\mu})(f(\mathbf{x}_u) - \boldsymbol{\mu})^T$ and $\mathbf{S}_B = \sum_{c\in\{1,\dots,N\}} M_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T$, where $\boldsymbol{\mu}$ is the mean of all embedded features associated with the pseudo-labeled set, M_c is the number of instances belonging to class c, $\boldsymbol{\mu}_c$ is the mean vector of the embedded features belonging to class c, and N is the number classes in the given few-shot task. Then, the *Fishers Criterion* (ψ) is defined as the ratio of the between-class scatter matrix to the scatter matrix:

$$\psi := \mathrm{tr}\{\bar{\mathbf{S}}^{-1}\mathbf{S}_B\} \tag{6}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace operation. To explain more, the eigen-vectors of matrix $\bar{\mathbf{S}}^{-1}\mathbf{S}_B$ compose the optimal space that maximises the between-class separability while minimising the within-class variability. The Fishers Criterion, calculated as the summation of the corresponding eigen-values, is regarded as a measure of the overall data separability.

Next, we evaluate the credibility of each pseudo-labeled instance $(\mathbf{x}_u, \hat{y}_u)$ by measuring its contribution to the overall discriminant power, i.e. to measure the difference of Fishers Criterion value when the instance is present and the instance is removed. Precisely, the influence of removing a specific instance on the discriminant power ψ is referred to as the *Instance Discriminant Analysis* (IDA):

$$d\psi_u := \mathrm{tr}\{\bar{\mathbf{S}}^{-1}\mathbf{S}_B\} - \mathrm{tr}\{[\bar{\mathbf{S}}\neg u]^{-1}\mathbf{S}_B\neg u\} \tag{7}$$

where $\bar{\mathbf{S}}\neg u$ and $\mathbf{S}_B\neg u$ are derived from the remaining data after removing instance u. $d\psi_u$ captures the reduction in the feature discriminant power caused by removing instance u, and it can be used as a metric for our sample selection process. Larger $d\psi_u$ indicates that the instance has greater (positive) impact to the data separability, thus its pseudo-label is more trustworthy and the instance should be selected to the augmented support set. We sort the pseudo-labeled examples in the descending order of their $d\psi_u$ value, and only select the top-ranking examples.

On the other hand, the exact computation of $d\psi_u$ can be expensive, since it requires multiple matrix inverse. In order to perform our IDA-based sample evaluation and selection more efficiently, we provide the following approximation of the $d\psi_u$, which can be computed without any matrix operations, with only inner-product and scaler operations. The proof is provided in the appendix.

Proposition 1 (Bound on $d\psi_u$)**.** *Instance Discriminant Analysis (IDA) $d\psi_u$ of a sample u is upper-bounded by:*

$$d\psi_u \leq \frac{\delta f(\mathbf{x}_u)^T f(\mathbf{x}_u)}{\rho(f(\mathbf{x}_u)^T f(\mathbf{x}_u) - \rho)} + \frac{H_{4,1/2}(\nu_u + f(\mathbf{x}_u)^T f(\mathbf{x}_u))}{\rho(M_u - 1)} + \frac{f(\mathbf{x}_u)^T f(\mathbf{x}_u)(\nu_u + f(\mathbf{x}_u)^T f(\mathbf{x}_u))}{\rho(f(\mathbf{x}_u)^T f(\mathbf{x}_u) - \rho)(M_u - 1)} \tag{8}$$

where $\delta = \sum_{c\in\{1,\dots,N\}} M_c \boldsymbol{\mu}_c^T \boldsymbol{\mu}_c$; $\rho > 0$ is the ridge parameter; M_u is the number of examples sharing the same pseudo label as $\mathbf{x}_u$ in the dataset (including $\mathbf{x}_u$); $H_{4,1/2} = \sum_{k=1}^{4} k^{-1/2}$ is the generalized harmonic number; M_c is number of examples that have pseudo label equal to class c; $\boldsymbol{\mu}_c$ is the mean of class c; $\boldsymbol{\mu}_u$ is the mean of class that example $\mathbf{x}_u$ belongs to; and $\nu_u = M_u[(\boldsymbol{\mu}_u^T\boldsymbol{\mu}_u)^2 - 4(\boldsymbol{\mu}_u^T\boldsymbol{\mu}_u)(\boldsymbol{\mu}_u^T f(\boldsymbol{x}_u)) + 2(f(\boldsymbol{x}_u)^T f(\boldsymbol{x}_u))(\boldsymbol{\mu}_u^T\boldsymbol{\mu}_u) + 2(\boldsymbol{\mu}_u^T f(\boldsymbol{x}_u))^2]^{1/2}$.

In practice, we iteratively select the most trustworthy pseudo-labeled examples based on their IDA values to augment the support set as shown in Algorithm 1: The classifier is firstly trained with the initial support examples using our DM loss objective Eq.(5). Then, the trained classifier is used to infer the pseudo-labels of the unlabeled set. We employ the IDA method to select the most faithful ones into

Method	Setting	*mini*-ImageNet		CUB		CIFARFS		*tiered*-ImageNet	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
DSN [24]	In.	62.64	78.83	-	-	72.30	85.10	66.22	82.79
FEAT [32]		66.78	82.05	-	-	-	-	70.80	84.79
DeepEMD [34]		65.91	82.41	75.65	88.69	-	-	71.16	86.03
TransFinetune† [6]	Tran.	65.73	78.40	-	-	76.58	85.79	73.34	85.50
LaplacianShot† [36]		74.86	84.13	80.96	88.68	-	-	80.18	87.56
TIM† [4]		77.80	87.40	82.20	90.80	-	-	82.10	89.80
SIB† [11]		70.00	79.20	-	-	80.00	85.30	-	-
EPNet† [22]		70.74	84.34	87.75	94.03	-	-	78.50	88.36
ICI+LR [31]		66.80	79.26	88.06	92.53	73.97	84.13	80.79	87.92
BD-CSPN† [16]		70.31	81.89	87.45	91.74			78.74	86.92
LST [14]	Semi.	70.10	78.70	-	-	-	-	77.70	85.20
ICA+MSP† [15]		80.11	85.78	-	-	-	-	86.00	89.39
EPNet† [22]		79.22	88.05	-	-	-	-	83.69	89.34
ICI+LR [31]		71.41	81.12	91.11	92.98	78.07	84.76	85.44	89.12
Ours	Tran.	80.60	87.02	92.43	94.77	79.52	86.16	85.87	89.61
	Semi.	**83.34**	**88.17**	**93.51**	**95.44**	**82.16**	**87.26**	**87.12**	**90.54**

Table 2. Comparison of testing accuracy with previous state-of-the-art methods on four few-shot benchmark datasets. Our method outperforms previous state-of-the-art methods across all four few-shot classification benchmarks. 'In.' and 'Tran.' denotes inductive and transductive FSL, respectively. 'Semi.' denotes semi-supervised FSL. '-' denotes the results are not provided by the corresponding method. Methods with '†' use WRN28-10 as the backbone network.

the support set. The expanded support set is used to update the classifier based on Eq.(5) again. We iterate the above process to progressively enhance the classifier until the predicted pseudo-labels for the unlabeled set becomes stable.

5. Experiments

We evaluate on four widely used few-shot benchmark datasets: ***mini*-ImageNet** [29] consists of 100 classes, and we follow the split of base/novel classes as [20]; ***tiered*-ImageNet** contains 608 classes and we follow the spilt as [21]; **CUB** [30] is a fine-grained classification dataset, containing 200 classes and we follow the split as [5]; **CIFARFS** is a low-resolution few-shot dataset, containing 100 classes and we follow the split as [31].

Throughout the experiments, the hyperparameters of DM loss and IDA algorithm are kept fixed. Specifically, we use Guassian kernel with bandwith $\sigma = 0.5$ for the DM loss, and weight λ in Eq.(5) is set to 0.01. For the IDA algorithm, we select at most 5 samples per class at each iteration, until the pseudo-labeling process becomes stable. For training the softmax classifier, we use ADAM optimizer with 10^{-4} learning rate and run 1000 iterations for each task. Unless otherwise specified, we use WRN28-10 [33] as our main backbone for feature extraction, as it has been widely used by previous FSL works. Training the backbone network follows the same training procedure (without episodic training) as [4] on base classes. The models are trained for 90 epochs, with initial learning rate 0.1, divided by 10 at epochs 1/2 and 2/3, and batch-size 128. We employ standard data augmentation strategies, including random resized cropping, color jittering, and random horizontal flipping. All input images have resolution 84×84. To evaluate the testing performance, we randomly sample 600 few-shot tasks from the novel classes and report the averaged accuracy, following the same setup as [31].

5.1. Results of few-shot classification

Table 2 evaluates our method on the four benchmarks, under both transductive and semi-supervised setting.

Transductive FSL. In TFSL (denoted as Tran.), we have access to the query examples in the inference stage, thus we take the query set as the unlabeled set and utilize our proposed DM and IDA algorithms (i.e. no additional unlabeled set, but using the query set as the unlabeled set in Trans.) This transductive setup is used by all other compared methods as well. As shown in Table 2, our method outperforms previous state-of-the-art TFSL approaches across all datasets, especially on 1-shot tasks where the labeled support data is extremely limited. For example, on 1-shot *mini*-ImageNet, our method outperforms [4] by 2.8% accuracy.

Semi-Supervised FSL. We follow the SSFSL (denoted as Semi.) setup in [15, 22, 31], where each testing task has an additional unlabeled set consisting of unlabeled examples from the classes in the support set. The results are shown in Table 2. On *mini*-ImageNet and *tiered*-ImageNet datasets, we use 50 unlabeled examples per class in both 1-shot and

Method	*mini*-ImageNet→ CUB	
	1-shot	5-shot
MAML [7]	-	51.34
ProtoNet [25]	-	62.02
RelationNet [27]	-	57.71
Finetuning [5]	48.56	65.57
LaplacianShot § [11]	**55.46**	66.33
Ours (Tran.) §	**55.79**	**71.01**

Table 3. Results of testing accuracy for cross-domain FSL scenario. For a fair comparison, we use the same ResNet-18 backbone as compared methods. '§': denote transductive FSL methods. '-': results not reported.

Method	10-way		20-way	
	1-shot	5-shot	1-shot	5-shot
Baseline++ [5]	40.43	56.89	26.92	42.80
LEO [23]	45.26	64.36	31.42	50.48
MetaOpt [12]	44.83	64.49	31.50	51.25
S2M2$_R$ [18]	50.40	70.93	36.50	58.36
EPNet § [22]	53.70	72.17	38.55	59.01
BD-CSPN § [16]	51.58	69.35	36.00	55.23
Ours (Tran.) §	**60.05**	**75.93**	**41.47**	**61.91**

Table 4. Results of testing accuracy for higher-way scenario on the *mini*-ImageNet, where each testing task contains 10 or 20 unseen object categories. For a fair comparison, all methods are based on the WRN28-10 backbone. '§': denote transductive FSL methods.

5-shot scenarios. On CIFARFS and CUB datasets, we use 80 and 30 unlabeled examples per class, respectively. Compared with previous SSFSL method [14] that uses more than 100 unlabeled examples per class, our method still achieves better testing accuracy across all datasets with accuracy improvement ranging 5%∼13%. Compared with previous state-of-the-art [15], our method achieves 3.2% accuracy improves on 1-shot *mini*-ImageNet. Moreover, comparing our SSFSL results versus our TFSL results, we see that the additional unlabeled examples helps to further improve the testing accuracy.

Cross-domain FSL. [5] showed that many of meta-learning algorithms perform no better than the simplest fine-tuning baseline when there exists a domain-shift between the base dataset for training and the novel dataset for testing. We evaluate our method in this challenging cross-domain scenario, where we train the backbone network on the *mini*-ImageNet dataset while testing it on the few-shot tasks sampled from the CUB dataset. As shown in Table 3, our method outperforms previous meta-learning and transductive FSL methods, suggesting that our method is applicable to more realistic few-shot learning problems.

Higher-way testing scenario. We evaluate on more challenging 10-way and 20-way few-shot scenarios, where each task contains 10 or 20 unseen object categories but each class still only has 1 or 5 labeled examples. As shown in Table 4, compared with previous meta-learning or transductive FSL methods, our method achieves the highest accuracy on the difficult testing tasks with more object categories.

Loss	*mini*-ImageNet		*tiered*-ImageNet	
	1-shot	5-shot	1-shot	5-shot
Cross-entropy (CE) only	57.73	78.17	68.29	85.31
CE + Entropy [6]	65.73	78.40	73.34	85.50
CE + Mutual Info. [4]	71.54	83.92	77.02	87.57
CE + Dependency (Ours)	**75.80**	**85.26**	**82.42**	**89.12**

Table 5. Ablation study on the effectiveness of proposed dependency maximization loss. Results are testing accuracy. All methods use WRN28-10 as the backbone.

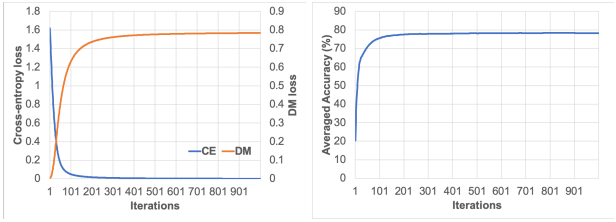


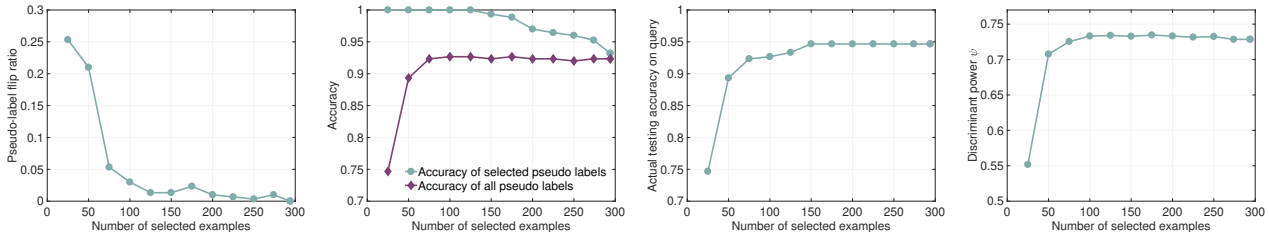
Figure 1. Convergence plot while we use Eq.(5) to train the classifer on *mini*-ImageNet 1-shot tasks. Left: cross-entropy loss and DM loss versus iterations. Right: accuracy of the classifier trained by the joint cross-entropy loss and DM loss versus iterations.

5.2. Ablation studies

We investigate the effectiveness of various components our method, namely the DM loss, the IDA method, and the self-attention based feature preprocessing in the following ablation studies.

Effectiveness of DM. To validate the effectiveness of DM loss, we compare it with several recently proposed label-free loss functions utilizing unsupervised information in query data for transductive FSL. Results are reported in Table 5. [6] proposes to minimize the conditional entropy of the label predictions over query data; [4] proposes to maximize the weighted mutual information over query data. Nevertheless, we can observe that incorporating our DM loss consistently outperform other types of transductive learning on both *mini*-ImageNet and *tiered*-ImageNet. This suggests that maximizing the dependency between query feature and the label predictions can effectively improve the generalization performance. Furthermore, Figure 1 shows the convergence plot for our DM method on 1-shot *mini*-ImageNet tasks. During training, DM value increases monotonically at each iteration and converges well.

Effectiveness of IDA. To further validate the effectiveness of IDA, we compare it with other metrics for eval-



(a) (b) (c) (d)

Figure 2. (a) The flip ratio of the pseudo-labels between two consecutive iterations in our IDA method. (b) The accuracy of the pseudo-labels w.r.t. the ground-truth labels. (c) The actual testing accuracy on the query versus number of selected pseudo-labeled examples. (d) The discriminant power ψ (Eq.(6)) of the pseudo-labeled set over iterations.

Metric	Transductive		Semi-supervised	
	1-shot	5-shot	1-shot	5-shot
No selection	56.06	75.43	56.06	75.43
Random	59.01	76.38	59.46	76.58
Nearest Neighbour	63.24	77.63	63.10	77.75
Confidence	63.29	77.92	63.57	77.71
ICI [31]	65.32	78.30	64.60	77.96
IDA (Ours)	**67.17**	**80.00**	**67.36**	**80.18**

Table 6. Comparing IDA to other metrics for selecting the pseudo-labeled examples for self-training. Results are testing accuracy on *mini*-ImageNet. All methods use ResNet-12 as the backbone.

1-shot Accuracy	*mini*-ImageNet		*tiered*-ImageNet	
	Tran.	Semi.	Tran.	Semi.
w/o SA-FP	79.17	82.46	84.10	**87.12**
w/ SA-FP	**80.60**	**83.34**	**85.87**	86.71

Table 7. Ablation study on the effectiveness of the self-attention based feature preprocessing (SA-FP).

uating the credibility and selecting the pseudo-labeled examples under the transductive/semi-supervised FSL setting in Table 6. A naive strategy is to randomly select some pseudo-labeled examples into the augmented support set. Another strategy is to select high-confidence examples, i.e. selecting the pseudo labels whose largest class probability given by the classifier is above a certain threshold [14]. One can also leverage the nearest-neighbour strategy to select the examples based on their distance to the centroid of each class in the feature space. The last strategy we compare to is ICI [31], which selects pseudo-labels based on a linear regression hypothesis. Here, we assume 15 unlabeled examples for each class and select 5 examples per class by different metrics to retrain the classifier on *mini*-ImageNet. As shown, IDA outperforms other metrics in all settings. Specifically, IDA outperforms the ICI pseudo-label selection method, suggesting that our discriminant hypothesis can select more faithful pseudo-labeled examples thatn the linear regression hypothesis in [31]. We also provide visualizations for our iterative IDA method in Figure 2. We have the following observations: (1) the pseudo labels gradually stabilize without label flipping at the end; (2) IDA can select more trustworthy pseudo labels as the accuracy of the selected pseudo labels is always higher than the accuracy of overall pseudo labels; (3) the accuracy of overall pseudo-labels gradually improves, indicating that some wrongly labeled examples in previous iterations can be corrected and re-considered to select in later iterations; (4) the actual testing accuracy on query consistently improves as we attractively select more pseudo labeled examples.

Effectiveness of self-attention feature preprocessing. The self-attention feature preprocessing step is to drive similar features in the unlabeled set to cluster closer and to propagate the unlabeled feature information to the support features to better represent the data distribution of the task. In Table 7, we see that this preprocessing step improves the testing performance noticeably for both transductive and semi-supervised FSL on *mini*-ImageNet and for transdutive FSL on *tiered*-ImageNet.

6. Conclusion

Few-shot learning is a fundamental problem in modern AI research. In this paper, we propose a simple approach to exploit unlabeled data to improve the few-shot performance. We propose a dependency maximization loss based on the Hilbert-Schmidt norm of the cross-covariance operator, which maximizes the statistical dependency between the features of unlabeled data and their label predictions. The obtained model can be used to infer the pseudo-labels for the unlabeled data. We further propose an instance discriminant analysis to evaluate the quality of each pseudo-labeled example and only select the most faithful ones to augment the support set. Extensive experiments show that our method compares favourably with state-of-the-art methods on standard few-shot benchmarks, as well as on higher-way testing tasks and cross-domain FSL. In future work, we aim to provide a more theoretical analysis for our dependency maximization and discriminant-based sample selection; and we aim to generalize the method to other applications such as few-shot detection/segmentation.

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018. 2

[2] Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 1973. 4

[3] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. *CVPR*, 2020. 2

[4] Malik Boudiaf, Ziko Imtiaz Masud, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Transductive information maximization for few-shot learning. *NeurIPS*, 2020. 2, 3, 6, 7

[5] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *ICLR*, 2019. 1, 6, 7

[6] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *ICLR*, 2020. 2, 6, 7

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017. 2, 7

[8] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. *NeurIPS*, 2018. 2

[9] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. *ALT*, 2005. 4

[10] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *NeurIPS*, 2019. 2

[11] 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. *ICLR*, 2020. 2, 6, 7

[12] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. *CVPR*, 2019. 7

[13] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. *CVPR*, 2020. 2

[14] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. *NeurIPS*, 2019. 2, 6, 7, 8

[15] Moshe Lichtenstein, Prasanna Sattigeri, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Tafssl: Task-adaptive feature sub-space learning for few-shot classification. *ECCV*, 2020. 6, 7

[16] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. *ECCV*, 2020. 6, 7

[17] 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. 2

[18] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. *WACV*, 2020. 7

[19] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. *ICML*, 2007. 1, 3

[20] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *ICLR*, 2017. 6

[21] 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. *ICLR*, 2018. 2, 6

[22] Pau Rodríguez, Issam Laradji, Alexandre Drouin, and Alexandre Lacoste. Embedding propagation: Smoother manifold for few-shot classification. *ECCV*, 2020. 2, 6, 7

[23] 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*, 2019. 2, 7

[24] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. *CVPR*, 2020. 2, 6

[25] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *NeurIPS*, 2017. 2, 7

[26] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. *CVPR*, 2019. 2

[27] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. *CVPR*, 2018. 2, 7

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[29] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *NeurIPS*, 2016. 2, 6

[30] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[31] Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu. Instance credibility inference for few-shot learning. *CVPR*, 2020. 2, 6, 8

[32] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. *CVPR*, 2020. 2, 6

[33] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *BMVC*, 2016. 6

[34] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. *CVPR*, 2020. 2, 6

[35] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. *NeurIPS*, 2018. 2

[36] Imtiaz Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. *ICML*, 2020. 6