

Transductive Episodic-Wise Adaptive Metric for Few-Shot Learning

Limeng Qiao^{1,4}, Yemin Shi^{2,4}, Jia Li^{3,4*}, Yaowei Wang⁴, Tiejun Huang^{2,4} and Yonghong Tian^{2,4*},

¹ Center for Data Science, AAIS, Peking University

² National Engineering Laboratory for Video Technology, School of EE&CS, Peking University

³ State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University

⁴ Peng Cheng Laboratory, Shenzhen, China

Abstract

Few-shot learning, which aims at extracting new concepts rapidly from extremely few examples of novel classes, has been featured into the meta-learning paradigm recently. Yet, the key challenge of how to learn a generalizable classifier with the capability of adapting to specific tasks with severely limited data still remains in this domain. To this end, we propose a Transductive Episodic-wise Adaptive Metric (TEAM) framework for few-shot learning, by integrating the meta-learning paradigm with both deep metric learning and transductive inference. With exploring the pairwise constraints and regularization prior within each task, we explicitly formulate the adaptation procedure into a standard semi-definite programming problem. By solving the problem with its closed-form solution on the fly with the setup of transduction, our approach efficiently tailors an episodic-wise metric for each task to adapt all features from a shared task-agnostic embedding space into a more discriminative task-specific metric space. Moreover, we further leverage an attention-based bi-directional similarity strategy for extracting the more robust relationship between queries and prototypes. Extensive experiments on three benchmark datasets show that our framework is superior to other existing approaches and achieves the state-of-the-art performance in the few-shot literature.

1. Introduction

Deep neural networks have achieved great success on many practical applications recently and even have surpassed humans in image recognition domain [8, 15, 28]. However, these successes largely benefit from tens of thousands of labeled data, enormous number of parameters and sophisticated training strategies. In terms of the low-data scenarios, such as medical images classification and ma-

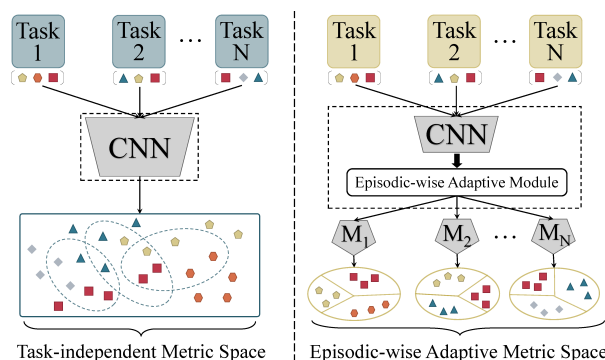


Figure 1. The left part shows a general framework of previous approaches based on meta learning. The model embeds all samples into a task-independent metric space with a shared CNN. However, considering the distinctive information within each task, we further learn an episodic-wise adaptive metric for classification. The right part illustrates the motivation of our approach briefly.

rine biological recognition, deep neural networks will be over-fitting and severely collapse because only rare labeled samples are provided for training. To this end, by exploring the exciting idea that tries to learn new concepts rapidly and generalize well with extremely few examples, or even single, few-shot learning has attracted significant research interest recently [5, 19, 24, 27, 29, 35, 38].

Concretely, we focus on the case of few-shot classification with the meta-learning paradigm, which leverages a series of independent and identically distributed few-shot tasks to learn a desired classifier at training time and applies the model directly to non-overlap unseen target classification problem during testing phase. The so-called episodic-training strategy, which has been widely developed in many previous work [29, 35], keeps the consistency between the training and real test scenario and improve the generalization performances. In each task (or episode), only a handful of labeled examples per classes (the support set) are provided to the classifier for training and then plenty of unlabeled points (the query set) need to be assigned labels for prediction. Specifically, [29, 35, 38] jointed the meta-learning with metric learning as a feed-forward manner, with opti-

*Correspondence should be addressed to Yonghong Tian and Jia Li. (yhtian@pku.edu.cn and jiali@buaa.edu.cn).

mizing a shared distance metric across all tasks. Moreover, [5] proposed to learn a general initialization strategy and [17, 24] learned to update the parameters of learner directly with a higher-level meta-optimizer (e.g. LSTM).

While these approaches based on meta-learning paradigm have made significant advances in few-shot classification, they do suffer from two distinct limitations. One is that all examples from various different tasks are embedded into a task-independent metric space indiscriminately, see Fig. 1 (left). Namely, this assumption does not take the task-level information (meta-data) into account but example-level feature only, which neglects the specificity of different tasks. Actually, what is missing of this idea is an adaptive module that tailors the metric space for each task. The other is that most current methods follow the setup of inductive inference, that is, training the meta-learner with severely limited support data and predicting queries one by one in each task. Obviously, this process does not adequately consider the interaction between the support set and unlabeled test set and thus weakens the advantages of meta-learning.

To deal with the key challenge of how to learn a generalizable classifier with the capability of adapting to specific tasks with severely limited data, we propose a novel meta-learning framework for few-shot classification, named as *Transductive Episodic-wise Adaptive Metric* (TEAM), which efficiently tailors an episodic-wise metric space with applying the idea of transductive inference. Specifically, we put forward not only to learn a task-agnostic instance embedding model end-to-end over a pool of few-shot tasks as a meta-learning manner, but also constructs a task-specific distance metric explicitly for each task with the distinctive information, such as the pairwise constraints and regularization prior, see Fig. 1 (right). Furthermore, we formulate the optimization process for task-specific metric into a standard semi-definite programming (SDP) problem [2] and finally acquire the closed-form solution by solving the SDP problem on the fly. Hereafter all features generated by the task-agnostic model are adapted into task-specific metric spaces where samples from the same class are closer and different classes are farther apart. According to the transformed embeddings, we then perform a novel attention-based bi-directional similarity strategy to compute more robust relationship between each unlabeled query and class prototypes, which further improves the performance of our approach. In addition, by utilizing the convex combination of all samples within each task to construct auxiliary training tasks, we propose a task-level data augmentation technique for boosting the generalization of the embedding model. The whole framework is illustrated in Fig. 2 in details.

The main contribution is summarized as threefold. (1) We propose a general meta-learning framework of applying the transductive inference on formulating the adaptation procedure into a SDP problem and tailoring the episodic-

wise metric in each task for few-shot learning, which can also be directly extended to other existing methods and even semi-supervised learning. (2) We identify a novel bi-directional similarity strategy for extracting the more robust relationship between queries and prototypes. (3) The experimental results on three benchmark datasets show that our framework is superior to other state-of-the-art approaches.

2. Related Work

Meta-learning in Few-shot Learning. Meta-learning, or learning to learn [31], is the technique of observing how different machine learning approaches perform on a wide range of tasks rather than batches of data points, and then starting from this experience, or meta-data, to learn new tasks much faster and obtain higher performance. In recent few-shot learning literature, more and more approaches follow the idea of meta-learning to alleviate over-fitting. Meta-LSTM [24] aims to learn efficient parameter updating rules for training neural network (learner) with a LSTM-based meta-learner. MAML [5], on the other hand, tries to learn a good model parameter initialization strategy that generalizes better to similar tasks. Similarly, Reptile [21] is an approximation of MAML that executes stochastic gradient descent with a first-order form. Meta-SGD [17] goes further in meta-learning by arguing to learn the weights initialization, gradient update direction and learning rate within a single step. However, these approaches mentioned above suffer from the issue of fine-tuning. In contrast, after optimizing a meta-learner over a series of tasks, our approach solves unseen target tasks in an feed-forward manner without any further model updates and just optimize an episodic-wise distance metric efficiently for each task, thus avoiding gradient computation and serious overfitting.

Distance Metric Learning Approaches. Another category of approach focus on obtaining a generalizable embedding model to transform all samples into a common metric space where can perform simple classifiers such as nearest neighbor directly. Matching Network [35] integrates metric learning with meta-learning for the first time by training a learnable nearest neighbor classifier with deep neural networks. Prototypical Network [29] utilizes class prototype representations to assign labels for query points and formulates the final loss function with euclidean distance directly. Later, this work is extended to semi-supervised few-shot scenario by [25], unlabeled data are used to refine the class prototypes. Relation Network [38] trains an auxiliary network to compute the similarity score between each query and the support set, which is equivalent to further learn a non-linear metric. These approaches assume all samples are embedded into a task-agnostic metric space. Instead, our framework intends to emphasize the specificity among different tasks, that is, samples from different tasks should be embedded into a more discriminative task-specific space.

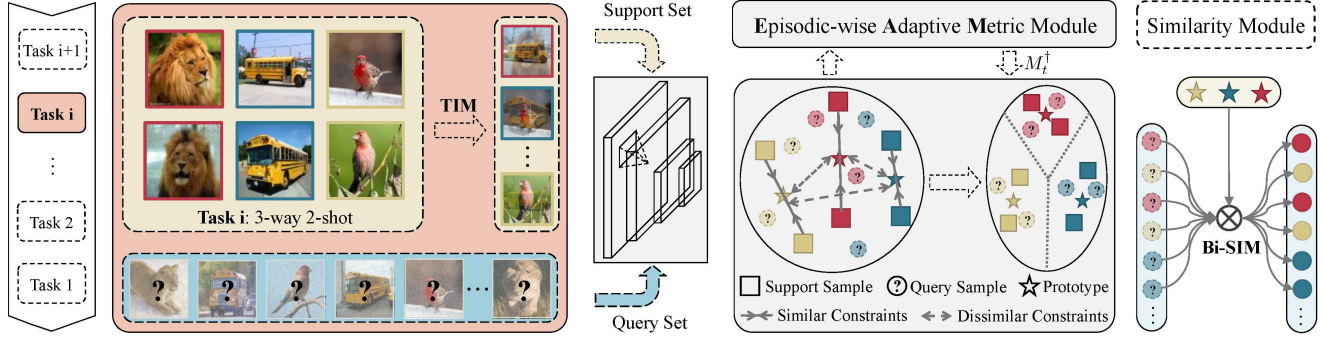


Figure 2. The architecture of Transductive Episodic-wise Adaptive Metric (TEAM) framework for few-shot learning. The whole training data is reorganized into a pool of few-shot tasks (meta-training set) firstly and then the task-level data augmentation technique (TIM) performs the convex combination between all samples in each task. After that, the augmented episodic data is fed into the CNN and the EAM module constructs a task-specific distance metric M_t^\dagger explicitly for each task on the fly with the pairwise constraints and regularization prior. The symbol \otimes in Bi-SIM module indicates that each query is assigned a label using the bi-directional similarity.

Task Adaptation Approaches. The third family of approaches, which aims to mine the task adaptability with the meta-learning paradigm, is currently a hot research direction for few-shot learning. MT-Net [16] proposes that the meta-learner learns on each layers activation space and the task-specific learner performs gradient descent on a subspace, which is more sensitive to task identity. TADAM [22] proposes a Task Encoding Network to produce scaling and shift vectors for each layer weights, leading to a task-dependent metric space. Our approach is slightly similar to these methods in the sense that we all focus on the idea of task adaptation. However, without imposing auxiliary network or applying complicated training strategy, our proposed framework formulates the adaptation procedure into a standard SDP problem under the transductive inference setting, which is more efficient and convenient.

Transductive Inference. Transductive inference [34] follows the setting of generalizing from the training set to the test set directly and avoiding the intermediate problem of estimating a function. In data-scarce scenario, it can significantly improve the performance over inductive methods. Reptile [21], which implicitly shared information between all test samples via batch normalization [11], is the first work of applying the transductive setting in few-shot learning. TPN [18] models the transductive inference explicitly by learning a graph construction module, which propagates labels from labeled instances to unlabeled query points directly. Different from using example-wise parameters for the static Euclidean distance in [18], our approach explores the episodic-wise distance metric by integrating the small support set with the entire query set for transduction.

3. Transductive Episodic-wise Adaptive Metric

In this section, we describe the definition of few-shot learning problem (FSL) and then introduce the *Transductive Episodic-wise Adaptive Metric* (TEAM) in details.

3.1. Problem Formulation

Just like what is employed in various previous works [25, 29, 35, 38], we organize the learning procedure into the form of episodic paradigm, which gradually collects meta-knowledge across a pool of source tasks and performs adaptation on target tasks quickly. Under this setting, the ultimate goal of our algorithm is to train models with a large labeled dataset \mathcal{D}_{train} , which is composed of a set of seen classes \mathcal{C}_{train} , and apply the classifiers on a novel testing set \mathcal{D}_{test} with many unseen classes \mathcal{C}_{test} . Note that there are only a few labeled examples for each category in \mathcal{C}_{test} and $\mathcal{C}_{test} \cap \mathcal{C}_{train} = \emptyset$. In order to mimic the few-shot test scenario during the training procedure and take full advantage of the large quantities of labeled \mathcal{D}_{train} , we reorganize all examples in \mathcal{D}_{train} as a series of N -way K -shot tasks (or episodes). Concretely, N -way K -shot task is usually constructed by first selecting N classes from \mathcal{C}_{train} randomly and then generate a support set and a query set from the selected classes. The support set includes K samples per classes, termed as $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N \times K}$, and whilst the query set $\mathcal{Q} = \{(\hat{x}_1, \hat{y}_1), \dots, (\hat{x}_M, \hat{y}_M)\}$ contains different samples from the same label space with \mathcal{S} . In each episode, we train the learner with small labeled support set \mathcal{S} and minimize the loss on the large query set \mathcal{Q} . After training episode by episode until convergence, the learned model can perform pretty well on novel few-shot tasks.

However, we argue that it is not ideal to apply the learned model to all target tasks directly without considering the specificity of them. Our approach proposes an episodic-wise metric construction module to transform the task-agnostic embeddings into a task-specific metric space, namely episodic adaptation. Moreover, in order to mitigate the data-scarce problem of support set and construct a more generalizable task-adaptive metric, we follow the paradigm of transductive inference and consider the query set as a whole for prediction, instead of one by one.

Specifically, our few-shot classification framework (see Fig. 2) is composed of three modules: (1) learning a task-agnostic feature extractor to embed raw inputs into a shared embedding space, this procedure includes a loss function to drive the parameters update and a novel task-level augmentation strategy to boost the generalization, (2) tailoring an episodic-wise adaptive metric for each task by solving a standard SDP problem efficiently and (3) performing a novel bi-directional similarity strategy in the task-specific space to assign label for each query. It is worth noting that the latter two modules utilize the setting of transduction. We describe the details of each module in the following sections.

3.2. Task-agnostic Feature Extractor

Learning Embedding Function. Our approach first employs an embedding function f_θ to extract feature of an instance x , where $f_\theta(x)$ refers to the embedding of x and θ indicates the parameters of deep model. Given the few-shot task sequence \mathcal{T} from \mathcal{D}_{train} , we train the feature extractor f_θ episode by episode with minimizing the negative log-probability of true label for each sample via SGD as:

$$\underset{f_\theta}{\operatorname{argmin}} \sum_{\mathcal{T}_i \sim \mathcal{T}} \sum_{(x,y) \in \operatorname{TIM}(\mathcal{T}_i)} -\log p(y | f_\theta(x, M_{\mathcal{T}_i})) \quad (1)$$

where \mathcal{T}_i is a few-shot episode sampled from task sequence randomly, $\operatorname{TIM}(\cdot)$ is the novel task-level data augmentation operation and $M_{\mathcal{T}_i}$ stands for the episodic-wise adaptation metric of \mathcal{T}_i . Until convergence, the optimal embedding model f_θ^* is applied directly to unseen target tasks which are sampled from \mathcal{D}_{test} instead of \mathcal{D}_{train} .

Task Internal Mixing Augmentation Strategy. Recently, some data augmentation techniques, such as flipping, rotation or distorting the inputs, follow the learning principle named Vicinal Risk Minimization (VRM) [3] to improve the generalization performance of deep neural networks. Inspired by [10, 39], we further propose a task-level data augmentation technique which is termed as Task Internal Mixing (TIM), performing the convex combination between all support samples in each task to synthesize new episodes. To be concrete, for each instance (x_i, y_i) in a source task, we randomly select another sample (x_j, y_j) from the same task and synthesize new training examples (\tilde{x}, \tilde{y}) as follows:

$$\tilde{x} = \omega \cdot x_i + (1 - \omega) \cdot x_j, \quad \tilde{y} = y_i \quad (2)$$

where $\omega \sim U(l, h)$ and $0.5 \leq l < h \leq 1.0$. Note x_i and x_j are just the raw input tensors instead of features. Then we handle each instance a few times with Eq. (2) to form a virtual task (i.e. $\operatorname{TIM}(\mathcal{T}_i)$ in Eq. (1)) for training. In essence, TIM extends the source task distribution by incorporating the prior that if two samples are similar to each other in the original pixel space, then they are likely to be closer in the feature space. As such, x_i and \tilde{x} are more similar to each other than x_j and \tilde{x} in Eq. (2) because of $\omega > 0.5$, which leads to the synthetic label \tilde{y} should be y_i instead of y_j .

3.3. Episodic-wise Adaptive Metric

Distance with Metric M_t . Given two embeddings x_i and x_j in vector space $\mathcal{V} \in \mathbb{R}^d$, we denote the distance between them with the metric M_t as follows:

$$d_{M_t}(x_i, x_j) = \sqrt{\operatorname{tr}\langle M_t(x_i - x_j)(x_i - x_j)^T \rangle} \quad (3)$$

where $\operatorname{tr}\langle \cdot \rangle$ stands for the trace operator and M_t is a symmetric positive semi-definite matrix, which ensures that d_{M_t} satisfies the properties of a pseudo-distance [1]. In general, the matrix M_t parameterizes a family of Mahalanobis distances in the vector space \mathcal{V} . In particular, the d_{M_t} in Eq. (3) will degenerate into the popular Euclidean distance if we set $M_t = I$, that is, assuming all features are equally scaled and equally relevant [32]. Inspired by these observations, we proposed to explicitly construct episodic-wise adaptive metric by leveraging the specific information of each task, such as the pairwise constraints and the regularization prior.

Pair-constrained Loss. Given a few-shot task, our goal is to minimize the mean of distances between all similar sample pairs (must-link constraints, denoted with \mathcal{M}) while keeping the mean of distances between all dissimilar sample pairs (cannot-link constraints, denoted with \mathcal{C}) larger than 1 at meanwhile. Motivated by the above idea, we formulate the *Min-Max* principle as a convex optimization problem in Eq. (4) by adopting the square distance in terms of its effectiveness and efficiency.

$$\begin{aligned} \min_{M_t \succeq 0} \quad & \frac{1}{|\mathcal{M}|} \sum_{(x_i, x_j) \in \mathcal{M}} d_{M_t}^2(x_i, x_j) \\ \text{s.t.} \quad & \frac{1}{|\mathcal{C}|} \sum_{(x_i, x_j) \in \mathcal{C}} d_{M_t}^2(x_i, x_j) \geq 1 \end{aligned} \quad (4)$$

Based on Eq. (3) and the methods of lagrange multiplier, we rewrite the Eq. (4) into a pair-constrained loss function as:

$$\mathcal{L}_C(M_t | \tilde{\mathcal{M}}, \tilde{\mathcal{C}}) = \operatorname{tr}(M_t \cdot \tilde{\mathcal{M}}) - \lambda \cdot \operatorname{tr}(M_t \cdot \tilde{\mathcal{C}}) \quad (5)$$

where λ is the multiplier, $\tilde{\mathcal{M}}$ and $\tilde{\mathcal{C}}$ have the following form:

$$\begin{aligned} \tilde{\mathcal{M}} &= \frac{1}{|\mathcal{M}|} \sum_{(x_i, x_j) \in \mathcal{M}} (x_i - x_j)(x_i - x_j)^T \\ \tilde{\mathcal{C}} &= \frac{1}{|\mathcal{C}|} \sum_{(x_i, x_j) \in \mathcal{C}} (x_i - x_j)(x_i - x_j)^T \end{aligned} \quad (6)$$

Furthermore, given a N -way K -shot task with support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{NK}$ and its query set \mathcal{Q} , we first shrink the support set \mathcal{S} to a prototype set $\mathcal{P} = \{p_c\}_{c=1}^N$ by:

$$p_c = \frac{1}{|\mathcal{S}_c|} \sum_{(x_i, y_i) \in \mathcal{S}_c} x_i \quad (7)$$

where \mathcal{S}_c is the subset which contains the samples with the same label c in \mathcal{S} . Then we define the similar and dissimilar constraints with $\mathcal{M} = \{x_i \in \mathcal{S}_c, x_j \in \mathcal{S}_c, i \neq j\} \cup \{x_i \in \mathcal{S}_c, x_j \in \mathcal{P}_c\} \cup \{x_i \in \mathcal{S}_c, x_j \in \mathcal{N}(x_i, k, \mathcal{Q})\}$ and $\mathcal{C} = \{x_i \in \mathcal{P}_c, x_j \in \mathcal{P}_{c'}, c \neq c'\} \cup \{x_i \in \mathcal{P}_c, x_j \in \mathcal{P}_T\}$ where \mathcal{P}_c is a prototype with label c in \mathcal{P} , $\mathcal{N}(x_i, k, \mathcal{Q})$ is a set of k nearest neighbors of x_i in the query set \mathcal{Q} and \mathcal{P}_T is the prototype set from the seen classes \mathcal{C}_{train} .

Regularization Loss. Without imposing any restrictions or prior information into the Eq. (5), we have d^2 parameters to be optimized for each task due to $M_t \in \mathbb{R}^{d \times d}$, while we can only construct a few pairwise constraints in few-shot scenario. From the perspective of machine learning theory, this inconsistency will lead to severe over-fitting of our models. To this end, we propose the second principle to regularize the episodic-wise metric M_t to be close to a given metric M_0 , which is associated with the prior across all few-shot tasks. Specifically, we try to minimize the Bregman divergence $D_\Phi(M_t \| M_0) = \Phi(M_t) - \Phi(M_0) - \langle \nabla \Phi(M_0), M_t - M_0 \rangle$ between M_t and M_0 with the log-determinant function $\Phi(M) = -\log \det(M)$, which is a strict convex, continuously differentiable divergence function. Then we formulate the regularization loss function $\mathcal{L}_R(M_t | M_0)$ as:

$$\mathcal{L}_R(M_t | M_0) = \text{tr}\langle M_0^{-1} M_t \rangle - \log \det(M_t) \quad (8)$$

where $\text{tr}\langle \cdot \rangle$ means the trace operator on matrix and Eq. (8) ignores the constant term regarding M_0 . More precisely, with the information-theory, optimizing \mathcal{L}_R is equivalent to minimizing the KL divergence between two multivariate Gaussian distributions parameterized by M_t and M_0 .

Episodic-wise Adaptive Metric. By integrating two principles mentioned above, we formulate a novel episodic-wise adaptive metric (EAM) loss function for each task as:

$$\min_{M_t \succ 0} \mathcal{L}_R(M_t | M_0) + \gamma \cdot \mathcal{L}_C(M_t | \tilde{\mathcal{M}}, \tilde{\mathcal{C}}) \quad (9)$$

where \mathcal{L}_C and \mathcal{L}_R are defined in Eq. (5) and Eq. (8) respectively, γ is a positive trade-off parameter. In general, minimizing Eq. (9) with SGD-based optimizers or other existing convex optimization solvers [30] will produce a local optimal solution for each task. However, due to the high time complexity of SDP solvers and requiring too many iterations with SGD-based optimizers, this sub-optimization procedure within each task will lead to an inefficient learning process. Here, we propose a faster and more efficient approach to construct the episodic-wise metric for each task. Concretely, we first reformulate the Eq. (9) as follows:

$$\min_{M_t \succ 0} \text{tr}\langle M_t \cdot (\mathcal{M}_0^{-1} + \gamma \cdot \tilde{\mathcal{M}} - \gamma \lambda \cdot \tilde{\mathcal{C}}) \rangle - \log \det(M_t) \quad (10)$$

Based on the Lemma (1), we get the optimal solution M_t^* by letting $\mathcal{Y} = \mathcal{M}_0^{-1} + \gamma \cdot \tilde{\mathcal{M}} - \gamma \lambda \cdot \tilde{\mathcal{C}}$ and assuming $\mathcal{Y} \succ 0$,

$$M_t^* = (\mathcal{M}_0^{-1} + \gamma \cdot \tilde{\mathcal{M}} - \gamma \lambda \cdot \tilde{\mathcal{C}})^{-1} \quad (11)$$

And the assumption of $\mathcal{Y} \succ 0$ can always hold if we pick up a positive definite matrix as prior metric M_0 . In addition to considering the pairwise constraints and regularization prior, we further introduce the feature correlation information into the final metric M_t^\dagger with the task covariance matrix $\Sigma_t = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$, which results in:

$$M_t^\dagger = (\mathcal{M}_0^{-1} + \gamma \cdot \tilde{\mathcal{M}} - \gamma \lambda \cdot \tilde{\mathcal{C}})^{-1} + \alpha \cdot \Sigma_t \quad (12)$$

where α, λ, γ are positive trade-off parameters and $\tilde{\mathcal{M}}, \tilde{\mathcal{C}}$ are from Eq. (6). Moreover, with the insights of trans-

duction, we calculate the task covariance matrix Σ_t with both the support set and query set in each episode. Obviously, the Eq. (12) only involves simple matrix operations, such as inversion and transpose, which is more efficient than SGD-based optimizers and naive SDP solvers. In addition, as a symmetric positive definite matrix, another insights into the nature of the learned episodic-wise metric is an adaptive linear projection layer by expressing M_t^\dagger as $L_t^T L_t$, then the $d_{M_t}(x_i, x_j)$ in Eq.(3) can be formulated as $\sqrt{\text{tr}\langle (L_t x_i - L_t x_j)^T (L_t x_i - L_t x_j) \rangle}$, where $L_t \in \mathbb{R}^{r \times d}$ is a task-specific transformation matrix.

Lemma 1 Let \mathcal{X}, \mathcal{Y} be two symmetric positive-definite matrices of the same size, then the function $f(\mathcal{X}) = \text{tr}(\mathcal{X}\mathcal{Y}) - \log \det(\mathcal{X})$ is minimized uniquely by: $\mathcal{X}^* = \mathcal{Y}^{-1}$.

Proof. See supplementary material for more details.

3.4. Bi-directional Similarity

Assuming all samples have been transformed into a task-specific embedding space with the learned feature extractor f_ϕ and the episodic-wise adaptive metric M_t^\dagger in Eq. (12), we then perform a novel bi-directional similarity strategy (*Bi-SIM*) to calculate the probability that each query belongs to each category. In detail, after shrinking the support set \mathcal{S} to a prototype set with Eq.(7), we formulate the *positive-direction* similarity $s_{i \rightarrow c}$ between query x_i and each prototype p_c with the softmax function:

$$s_{i \rightarrow c} = \frac{\exp(-d_{M_t}(x_i, p_c))}{\sum_{c'} \exp(-d_{M_t}(x_i, p_{c'}))} \quad (13)$$

Most previous methods used this similarity as the final probability of the query x_i belonging to each category. However, taking the entire query set into account with transductive inference, we further compute the probability of prototype p_c belonging to each query x_i with the following equation:

$$s_{c \rightarrow i} = \frac{\exp(-d_{M_t}(p_c, x_i))}{\sum_{i'} \exp(-d_{M_t}(p_c, x_{i'}))} \quad (14)$$

We termed $s_{c \rightarrow i}$ as the *negative-direction* similarity which can be interpreted as an attention-based weight of the prototype p_c over the whole query set \mathcal{Q} . At last, we perform the product of $s_{i \rightarrow c}$ and $s_{c \rightarrow i}$ as the final bi-directional similarity (denoted with *Bi-Sim*) between the query x_i and the prototype p_c , i.e. $s_{i \leftrightarrow c} = s_{c \rightarrow i} \cdot s_{i \rightarrow c}$. Essentially, the basic idea behind the *Bi-Sim* strategy is that if a query is similar to one prototype and the prototype is also similar to the query, then we argue that they are more matching with each other. Without increasing any computational burden or requiring any human interaction, our proposed strategy calculates more robust similarity efficiently.

4. Experiments

In this section, we detail our experimental setting and compare TEAM with state-of-the-art approaches on three challenging datasets, i.e. *miniImageNet* [35], *Cifar-100* [14]

and CUB [36], which are widely used as few-shot classification benchmarks in the literature.

4.1. Datasets

miniImageNet. The *miniImageNet* dataset is the most popular benchmark in few-shot learning community, which is proposed by [35] originally. This dataset is composed of 100 classes selected from ImageNet [15] randomly, and each class has 600 images, which are resized to 84×84 pixels for fast training and inference. Note that we follow the setup provided by [24] which splits the total 100 classes into 64 classes, 16 classes and 20 classes for training, validation and evaluation respectively. The validation set is only used for tracking model generalization in all experiments.

Cifar-100. The *Cifar-100* [14] is a simple dataset for image classification and consists of 100 categories, each having 600 RGB images (32×32). We further split the whole dataset into 64 classes as seen categories for training, 16 and 20 classes for validation and testing respectively [40]. Compared with *miniImageNet*, *Cifar-100* keeps the simplicity of dataset and decreases the inference complexity.

CUB. CUB [36], which is a benchmark dataset for fine-grain classification initially, is composed of 11788 images over 200 birds classes. Follow the same partition as [9], we use 100 classes for training, and another two 50 classes as unseen classes for validation and evaluation. And all images are cropped with the provided bounding box [33].

4.2. Experimental Settings

Backbone Networks. For fair and comprehensive comparison with previous baselines, we employ two backbone networks as our embedding function. 1) A four-layers convolution network (ConvNet) and 2) A standard deep residual network (ResNet-18) are widely adopted in the few-shot learning literature. Specifically, the ConvNet contains 4 repeated convolutional blocks, where each block is composed of a convolution layer with 64 filters (3×3 kernel), a batch normalization layer [11], a ReLU non-linearity and a max-pooling layer with size 2. In addition, we empirically add a global average pooling layer as last for accelerating the convergence of the model and reducing the dimensionality of the features. All inputs are resized to $84 \times 84 \times 3$ uniformly and the final output dimension is 256 for each image accordingly. For ResNet, we utilize the standard architecture proposed by [8] and remove the last fc-layer for reducing parameters. Furthermore, all inputs are resized to $224 \times 224 \times 3$ like many previous works. After the last average pooling layer, it leads to a 512 vector for each image.

Training Strategy. All backbone networks are optimized via SGD by Adam [13] end-to-end on DGX-1. Follow the strategy in [23, 26], we pre-train the ConvNet to classify all seen classes and utilizing the optimal weights for model initialization, and we train ResNet from scratch for simplicity.

Moreover, We perform *TIM* strategy in all experiments and set $l = 0.5$ and $h = 1.0$ for $\mathcal{U}(l, h)$. Inspired by [10], we start *TIM* strategy after 5000 episodes and intermittently disable it during training procedure, that is, performing task mixing for Y episodes and then close it for the next Z episodes. We empirically set $Y = 4$ and $Z = 1$ in our experiments. We decay the learning rate half every 10000 episodes and set the patience of early stopping as 20000.

Parameter Setup in M_t^\dagger . In Eq. (9), we set the trade-off parameters as $\alpha = 2$, $\gamma = 0.2$, $\lambda = 0.01$ and the prior metric as $M_0 = I$ in all experiments. In general, the choice of M_0 is not fixed and has an important influence on generalization of the learned episodic-wise metric. However, based on the following two observation, we argue the identity matrix is a quite natural choice for M_0 . Firstly, learning from the Euclidean distance provides the most unbiased prior across all few-shot tasks, that is, assuming all features from the task-agnostic embedding space are equally scaled and equally relevant. Secondly, we observe that the optimal adaptive metric for each few-shot task is close to the identity matrix, which has been illustrated in Fig. 3 of the paper. Please zoom in the Fig. 3 or refer appendix for more details.

4.3. Few-shot Learning Results

To verify the effectiveness of our approach for few-shot classification, we compare the proposed TEAM framework with our re-implemented baseline (ProtoNet [29]) and many state-of-the-art methods in various setting on three benchmark datasets (*miniImageNet*, *Cifar-100* and CUB). For fair comparison with previous works, we focus on two popular few-shot learning settings, namely 5-way 1-shot and 5-way 5-shot tasks, which both contain 15 queries per episode for validation. In addition to the above setup, we also experimente with the transductive setting in all datasets, where the model utilizes the entire query set in each task. Specifically, we consider two types of transduction in our experiments. 1) Transductive batch normalization [5, 21], which shares information between all test examples via batch normalization layer, denoted with *BN* in all tables and 2) explicit transduction, which is first introduced into the few-shot learning by [18]. Moreover, to make the evaluation more convincing, we report the final mean accuracy over 1000 test trails for all experiments and present 95% confidence intervals of all results in our supplementary material. Please refer our appendix for more complete results.

Results on *miniImageNet*. Experimental results on *miniImageNet* are shown in Table 1, where we can see that our model achieves state-of-the-art performance with ConvNet backbone and competitive results with ResNet architecture. We re-implement ProtoNet as our baseline with the simple pre-train strategy proposed by [23], and achieve better performance than previously reported ones in [29]. Taking ConvNet as an example, we get 51.68% and 68.71% for 5-

Table 1. Few-shot classification accuracy on *miniImageNet*. All results are averaged over 1000 test tasks which are randomly selected from the testing set. Tran: different type of transduction.

Model	Tran.	5-Way 1-Shot		5-Way 5-Shot	
		ConvNet	ResNet	ConvNet	ResNet
MatchNet [35]	No	43.56	-	55.31	-
MAML [5]	BN	48.70	-	63.10	-
MAML+ [18]	Yes	50.83	-	66.19	-
Reptile [21]	BN	49.97	-	65.99	-
ProtoNet [29]	No	49.42	-	68.20	-
GNN [6]	No	50.33	-	64.02	-
RelationNet [38]	BN	50.44	-	65.32	-
PFA [23]	No	54.53	59.60	67.87	73.74
TADAM [22]	No	-	58.50	-	76.70
adaResNet [20]	No	-	56.88	-	71.94
LEO [26]	No	-	60.06	-	75.72
TPN [18]	Yes	55.51	59.46	69.86	75.65
Baseline (Ours)	No	51.68	55.25	68.71	70.58
TEAM (Ours)	Yes	56.57	60.07	72.04	75.90

way 1-shot and 5-way 5-shot respectively, which are slightly better than 49.42% and 68.20% in [29]. After applying the TEAM framework on the baseline, the performance has been further improved significantly. For example, the absolute promotion of TEAM over published state-of-the-art is 1.06% for 1-shot and 2.18% for 5-shot, over our baseline is 4.89% and 3.33% respectively. Note that the comparisons with PFA [23], LEO[26] and TADAM [22] are a bit unfair since we train the TEAM(ResNet) without any pre-train weights or including a classification objective, however, our model still achieves the best performance on 1-shot task.

Results on Cifar-100. Next we turn to the rich experiments evaluated on the Cifar-100, and all results are shown in Table 2 for comparison in detail. Note that all results of MatchNet [35], MAML [5] and DEML [40] in Table 2 refer to the reported performance in [40]. Compared with our baseline, whose accuracy is slightly higher than previous points, we notice that our TEAM(ConvNet) increases by 6.24 % on 1-shot tasks and 2.65 % on 5-shot tasks, which demonstrates the effectiveness of our approach.

Results on CUB. The CUB dataset [36] is proposed for fine-grain recognition initially and also widely utilized in current few-shot classification literature. From Table 3, we observe that the performance of our baseline is far better than the previous ProtoNet [29], that is because we preprocess all images with the provided bounding boxes [33] to reduce the impact of background on final performance. Comparing the TEAM with our re-implemented baseline, both ConvNet and ResNet backbones achieve outstanding performance over 1-shot and 5-shot tasks.

Further Analysis. All results which are summarized in the Table 1 - 3 indicate that our approach can consistently improve the few-shot learning performance on different datasets. This confirms that, under the setting of transductive inference, our model can efficiently tailor an episodic-wise adaptive metric for each task and perform a suitable

Table 2. Few-shot classification performance on Cifar-100.

Model	Tran.	5-Way 1-Shot		5-Way 5-Shot	
		ConvNet	ResNet	ConvNet	ResNet
MatchNet [35]	No	50.53	-	60.30	-
MAML [5]	BN	49.28	-	58.30	-
ProtoNet [29]	No	56.66	-	76.29	-
DEML [7]	No	-	61.62	-	77.94
Baseline (Ours)	No	57.83	66.30	76.40	80.46
TEAM (Ours)	Yes	64.07	70.43	79.05	81.25

Table 3. Few-shot classification performance on CUB.

Model	Tran.	5-Way 1-Shot		5-Way 5-Shot	
		ConvNet	ResNet	ConvNet	ResNet
MatchNet [35]	No	56.53	-	63.54	-
MAML [5]	BN	50.45	-	59.60	-
ProtoNet [29]	No	58.43	-	75.22	-
RelationNet [38]	BN	62.45	-	76.11	-
DEML [7]	No	-	66.95	-	77.11
TriNet [4]	No	-	69.61	-	84.10
Baseline (Ours)	No	69.39	74.55	82.78	85.98
TEAM (Ours)	Yes	75.71	80.16	86.04	87.17

similarity between all samples. Furthermore, we notice that the performance promotion of our approach in 1-shot scenario is more significant than that in 5-shot. This observation agrees with the nature of transduction [12, 18], where more training data are available, the less performance improvement will be. With regards to this, we then perform 5-way k -shot ($k=1, 3, 5, 7, 9$) experiments on *mini-ImageNet* and all results are shown in Table. 6. As the number of shots increases, we notice that our TEAM consistently outperforms our baseline with a large margin, but the performance improvement from TEAM decreases slightly, which further verifies the above analysis about transductive inference.

4.4. Ablation Study

Effectiveness of Different Modules. According to the previous analysis, the proposed TEAM framework is far superior to our baseline and becomes the new state-of-the-art approach in few-shot classification literature. As a necessary step to the ablation study, we first analyze how much each module (*TIM*, *EAM* and *Bi-Sim*) contributes to the ultimate performance. All results are shown in Table 4 in great details. Note that our baseline (ProtoNet) uses the ConvNet as backbone network and achieves higher performance on all three datasets (see the second row in Table 4) than previous performance because of the pre-train weights initialization. Furthermore, we perform various setting of TEAM framework as follows. 1) TEAM[‡] adds the *TIM* strategy into our baseline, 2) TEAM[†] utilizes the *TIM* strategy and episodic-wise adaptive metric (*EAM*) simultaneously and 3) TEAM combines all three modules together to get the ultimate performance. By comparing the second and third rows in Table 4, we observe that the *TIM* strategy can consistently improve the performance of all few-shot tasks. Then we

Table 4. Few-shot classification performance for ablation study. Proto (Ours): the baseline. TEAM[‡]: baseline+TIM. TEAM[†]: baseline+TIM+EAM. TEAM: baseline+TIM+EAM+Bi-SIM.

Model	miniImageNet		Cifar-100		CUB	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Proto [35]	49.42	68.20	56.66	76.29	58.43	75.22
Proto (Ours)	51.68	68.71	57.83	76.40	69.39	82.78
TEAM [‡]	52.97	70.45	59.56	77.65	70.27	84.68
TEAM [†]	55.35	71.59	62.76	78.80	75.06	86.06
TEAM	56.57	72.04	64.07	79.05	75.71	86.04

Table 5. *semi*-Supervised comparison on miniImageNet

Methods	5-way 1-shot	5-way 5-shot
Soft k -Means [25]	50.09 \pm 0.45	64.59 \pm 0.28
Soft k -Means+Cluster [25]	49.03 \pm 0.24	63.08 \pm 0.18
Masked Soft k -Means [25]	50.41 \pm 0.31	64.39 \pm 0.24
TPN-semi [18]	52.78 \pm 0.27	66.42 \pm 0.21
TEAM-semi (Ours)	54.81 \pm 0.59	68.92 \pm 0.38

further compare the TEAM[‡] and TEAM[†], where the only difference between them is whether using EAM module. Taking 1-shot task of miniImageNet as an example, TEAM[†] is 2.38% higher than TEAM[‡], which demonstrates that it is feasible to construct an episodic-wise adaptive metric for each task in few-shot learning. And the last row of Table 4 further shows the effectiveness of our entire framework.

Comparison with *semi*-Supervised Few-shot Learning.

From the perspective of unlabeled data, transductive inference is a special case of semi-supervised learning, that is, the former one directly uses test set as unlabeled data and the latter uses more auxiliary unlabeled data. As such, we propose a semi-supervised version of the TEAM framework, namely TEAM-semi, to compare it with other semi-supervised few-shot approaches. Specifically, following the labeled/unlabeled data split in [25], we use 40% and 60% in each class as labeled and unlabeled data respectively. Note that the support/query examples in each task are both randomly sampled from the labeled set only for fair comparison. All results, which are averaged over 10 random labeled/unlabeled partition of the training set, are reported in Table 5 in details. Compared with the previous state-of-the-art approaches TPN [18], our TEAM-semi framework increases by 2.03% and 2.50% for 1-shot/5-shot respectively, which verifies its ability to handle both supervised and *semi*-supervised few-shot classification.

Sparsity Nature of Episodic-wise Adaptive Metric. In this section we explore the sparsity nature of the episodic-wise adaptive metric in few-shot learning. Take a 5-way 5-shot task as an example, we set 15-queries in each class and exploit the classic LMNN algorithm [37] with all support and query samples to optimize an oracle metric, which ensures all examples in this task can be completely distinguished. Then we scale all elements of the metric into the region [0, 1] and visualize its heatmap in Fig. 3 (left). We

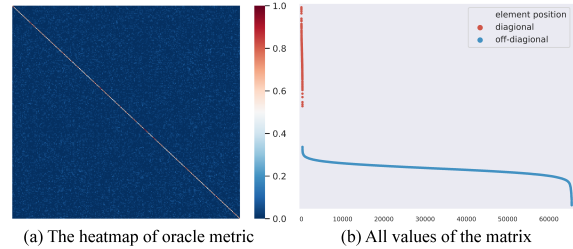


Figure 3. This figure illustrates the sparsity nature of the metric in few-shot learning. Left: The heatmap of the oracle metric (Please zoom in it for more details). Right: The values distribution in different position of the matrix (sorted by descending order).

Table 6. 5-way performance with various training/testing shots.

Methods	1-shot	3-shot	5-shot	7-shot	9-shot
Baseline (Ours)	51.68	63.87	68.71	71.28	73.35
TEAM (Ours)	56.57	67.64	72.04	73.47	75.04
Accuracy (+)	4.89	3.77	3.33	2.19	1.69

observe that diagonal elements always maintain larger values (close to red) than off-diagonal elements (close to blue). After reorganizing all values with numerical descending order in Fig. 3 (right), we further notice that there is a large value gap between the diagonal elements and off-diagonal elements. These practical observations indicate that, due to the low-data setup, we cannot have enough prior to find accurate correlations between all dimensions, except strong self-correlation in diagonal, which leads to the sparsity nature of episodic-wise adaptive metric. Moreover, from this practical viewpoint, we further verify that it is reasonable to set the identity matrix as prior metric M_0 in Eq. (12).

5. Conclusions

We have proposed Transductive Episodic-wise Adaptive Metric (TEAM) for few-shot learning, which is a simple and efficient framework based on meta-learning. It not only learns a shared embedding model across all tasks end-to-end but also further tailors an episodic-wise metric by taking more distinctive information within each task into account. Moreover, with using the entire query set at once for inference, we leverage a bi-directional similarity strategy for extracting more robust relationship between queries and prototypes. Our TEAM achieves the state-of-the-art performance on three few-shot benchmark datasets and is easily extended to *semi*-supervised version. The extensions of TEAM on other few-shot approaches could be future work.

Acknowledgement. This work is partially supported by grants from the National Key R&D Program of China under grant 2017YFB1002400, the National Natural Science Foundation of China under contract No.U1611461, No.61825101 and No.61672072, also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer, and Beijing Nova Program (Z181100006218063).

References

- [1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv:1306.6709*, 2013.
- [2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [3] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *NIPS*, pages 416–422, 2001.
- [4] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Semantic feature augmentation in few-shot learning. *arXiv:1804.05298*, 2018.
- [5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.
- [6] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *ICLR*, 2017.
- [7] Bharath Hariharan and Ross B Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, pages 3037–3046, 2017.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [9] Nathan Hilliard, Lawrence Phillips, Scott Howland, Artëm Yankov, Courtney D Corley, and Nathan O Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv:1802.04376*, 2018.
- [10] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv:1801.02929*, 2018.
- [11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [12] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.
- [13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [14] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
- [16] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *ICML*, 2018.
- [17] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few shot learning. *arXiv:1707.09835*, 2017.
- [18] 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. In *ICLR*, 2018.
- [19] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, 2017.
- [20] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, pages 3661–3670, 2018.
- [21] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR, abs/1803.02999*, 2, 2018.
- [22] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NIPS*, pages 719–729, 2018.
- [23] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, volume 2, 2017.
- [24] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *ICLR*, 2016.
- [25] 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. In *ICLR*, 2018.
- [26] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019.
- [27] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, pages 1842–1850, 2016.
- [28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017.
- [30] Jos F Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.
- [31] Sebastian Thrun and Lorian Pratt. *Learning to learn*. Springer Science & Business Media, 2012.
- [32] Lorenzo Torresani and Kuang-chih Lee. Large margin component analysis. In *NIPS*, pages 1385–1392, 2007.
- [33] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *NIPS*, pages 2255–2265, 2017.
- [34] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- [35] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016.
- [36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001*, 2011.
- [37] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2006.
- [38] Flood Sung Yongxin Yang, Li Zhang, Tao Xiang, Philip H-S Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

- [39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.
- [40] Fengwei Zhou, Bin Wu, and Zhenguo Li. Deep meta-learning: Learning to learn in the concept space. *arXiv:1802.03596*, 2018.