Ranking Distance Calibration for Cross-Domain Few-Shot Learning

Pan Li[1], Shaogang Gong[1], Chengjie Wang[2] and Yanwei Fu[3]
[1]Queen Mary University of London [2]Tencent Youtu Lab [3]Fudan University
{pan.li, s.gong}@qmul.ac.uk, jasoncjwang@tencent.com, yanweifu@fudan.edu.cn

Abstract

Recent progress in few-shot learning promotes a more realistic cross-domain setting, where the source and target datasets are in different domains. Due to the domain gap and disjoint label spaces between source and target datasets, their shared knowledge is extremely limited. This encourages us to explore more information in the target domain rather than to overly elaborate training strategies on the source domain as in many existing methods. Hence, we start from a generic representation pre-trained by a cross-entropy loss and a conventional distance-based classifier, along with an image retrieval view, to employ a re-ranking process to calibrate a target distance matrix by discovering the k-reciprocal neighbours within the task. Assuming the pre-trained representation is biased towards the source, we construct a non-linear subspace to minimise task-irrelevant features therewithin while keep more transferrable discriminative information by a hyperbolic tangent transformation. The calibrated distance in this target-aware non-linear subspace is complementary to that in the pre-trained representation. To impose such distance calibration information onto the pre-trained representation, a Kullback-Leibler divergence loss is employed to gradually guide the model towards the calibrated distance-based distribution. Extensive evaluations on eight target domains show that this target ranking calibration process can improve conventional distance-based classifiers in few-shot learning.

1. Introduction

Few-Shot Learning (FSL) promises to allow a machine to learn novel concepts from limited experience, *i.e.* few novel target data and data-rich source data. Typically, the defaulted FSL assumes that the source and target data is in the same domain, but belong to different classes. In practice, FSL is required to generalise to different target domains. Cross-Domain Few-Shot Learning (CD-FSL) [9, 10, 12, 20, 41] has been studied more recently. In CD-FSL, the target data not only has a different label space but also are from a different domain to the source data.

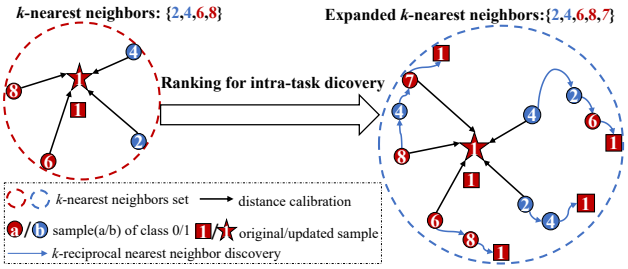


Figure 1. **An illustration of ranking distance calibration process in a FSL task**. The idea is to first discover likely positive samples (*e.g.* sample 7) for each instance (*e.g.* sample 1) and then to calibrate their pairwise distances. This is achieved by mining the reciprocal ranking relations for each instance retrieval task in the target domain so to expand the k-nearest neighbours set.

It is nontrivial to directly extend the general FSL approach to address the CD-FSL challenges. In fact, many promising FSL methods [8, 34, 36, 38] performed poorly in CD-FSL [10, 41]. The central idea of these general FSL methods is to transfer and generalise the visual representations learned from source data to target data. However, the significant visual domain gap between the source and target data in CD-FSL makes it fundamentally difficulty to learn a shared visual representation across different domains.

A few recent CD-FSL studies [20, 23, 41, 45] try to learn a generalisable feature extractor to improve model transferability, which is a popular idea in domain generalisation and domain adaptation [18, 43, 46, 52, 54] where the source and target domains share the same label space. Empirically, this approach shows some improvement on CD-FSL but it does not model any visual and label characteristics of the target domain and more importantly their cross-domain impact on the pre-trained source domain representation. We argue this cross-domain mapping between the source domain representation and its interpretion in the context of the target domain data characteristics is essential for effective CD-FSL. From a related perspective, other CD-FSL studies have considered fine-tuning the source domain feature representation from augmenting additional support data in the target domain, *e.g.* either explicitly augmenting the support data by adversarial training [45] and image transformations [10], or

implicitly augmenting the support data by training an auto-encoder [20]. However, these methods for CD-FSL are straightforward data-augmentation methods for increasing training data in target domain model fine-tuning, without considering how to quantify cross-domain relevance of the pre-trained source domain representation.

In this work, we consider an alternative approach with a new perspective to treat cross-domain few-shot learning as an image retrieval task. We wish to optimise model adaptation by leveraging target domain retrieval task context, that is, not only the labelled support data but also the unlabelled query data. To that end, we use a generic representation pre-trained by a cross entropy loss and a simple distance-based classifier as a baseline, then employ a k-reciprocal neighbour discovery (as in Fig. 1) and encoding process to calibrate pairwise distances between each unlabelled query image and its likely matches. Our idea is both orthogonal and complementary to other generalisable model learning methods [10, 20, 21]. It can be flexibly used in either fine-tuning or without fine-tuning based model learning.

Generally, the distance matrix for CD-FSL task contains many incorrect results as this distance is built on a potentially biased pre-trained source domain representational space. To calibrate this distance matrix towards the target domain so to reduce its bias to the source domain, we explore the re-ranking concept in the target domain by considering CD-FSL optimisation as re-ranking in a retrieval task given few-shots as anchor points. As in Fig. 1, re-ranking first computes a k-nearest neighbour ranking list. This is further expanded by discovering the k-reciprocal nearest neighbours in the target domain. The expanded ranking list is used for re-computing a Jaccard distance to measure the difference between the original ranking list and the expanded ranking list, achieving a more robust and accurate distance matrix. Critically, a pre-trained representation from source domain is biased and poor for generalisation cross-domains in CD-FSL. The reason is that conventional FSL methods assume implicitly linear transformations mostly between the source and target data as they are sampled from the same domain. This becomes invalid in CD-FSL with mostly nonlinear transformations across source and target domains. To address this problem, we propose a task-adaptive subspace mapping to minimise transferring task-irrelevant representational information from the source domain. In particular, we explore a hyperbolic tangent function to project the source domain representation to a non-linear space. Compared to the linear Euclidean space, this non-linear space performs a dimensionality reduction to optimise the retention of transferrable information from the source to the target domain. Moreover, we explore the idea of re-ranking to calibrate and align two distance matrices in two representational spaces between the original pre-trained source domain linear space and the new non-linear subspace. The calibrated matrices are combined to construct a single distance matrix for the target domain in CD-FSL. We call this Ranking Distance Calibration (RDC). To impose the above distance calibration into the representational space transform, we approximate the distance matrices by their corresponding distributions, and then a Kullback-Leibler (KL) divergence loss function is optimised for iteratively mapping the original distance distribution from the source domain towards the calibrated space. This provides an additional RDC Fine-Tuning (RDC-FT) model optimisation.

Our **contributions** from this work are three-fold: (1) To transform the biased distance matrix in the source domain representational space towards the target domain in CD-FSL, we use a re-ranking method to re-compute a Jaccard distance for distance calibration by discovering the reciprocal nearest neighbours within the task. We call this Ranking Distance Calibration. (2) We propose a non-linear subspace to shadow the pre-trained source domain representational space. This is designed to model any inherent non-linear transform in CD-FSL and used to facilitate the distance calibration process between the source and target domains. By modelling explicitly this nonlinearity, we formulate a more robust and generalisable Ranking Distance Calibration (RDC) model for CD-FSL. (3) We further impose RDC as a constraint to the model optimisation process. This is achieved by a RDC with Fine-Tuning (RDC-FT) for iteratively mapping the original source domain distance distribution to a calibrated target domain distance distribution for a more stable and improved CD-FSL.

We evaluated the proposed RDC and RDC-FT methods for CD-FSL on eight target domains. The results show that RDC can improve notably the conventional distance-based classifier, and RDC-FT can improve the representation for target domain to achieve competitive or better performance than the state-of-the-art CD-FSL models.

2. Related Work

Few-shot learning. The approaches for general FSL can be broadly divided into two categories: optimisation-based methods [8, 28, 32] which learn a generalisable model initialisation and then adapt the model on a novel task with limited labelled data, and metric learning methods [19, 36, 38, 51] that meta-learn a discriminative embedding space where the sample in novel task can be well-classified by a common or learned distance metric. Recently, some studies [3, 16, 39] show that a simple pre-training method followed by a fine-tuning stage can achieve competitive or better performance than the metric learning methods. This observation also seems to be true in CD-FSL [10].

Cross-domain few-shot learning. The problem of CD-FSL was preliminarily studied in FSL [3, 17, 39], then [10,

41] expanded this setup and proposed two benchmarks to train a model on a single source domain and then generalise it to other domains. Some CD-FSL studies [20,41,45] focus on learning a generalisable model from the source domain by explicit or implicit data augmentation. These approaches improve the model generalisation ability but easily result in ambiguous optimisation result since they ignore the adaption process for the target domain. To address the domain shift problem in a feature representation, CHEF [1] uses a fusion strategy for feature ensemble whilst ConFeSS [5] learns a mask to select relevant features in the target domain. Another methods [6, 9, 30] target to the adaptation on the target domain by leveraging additional unlabelled data [30], labelled data [9] or the base data [6]. In practice, the increased data can help model adaptation on the target domain but these information are not easy to obtain. In this work, we address the CD-FSL problem with an image retrieval view and mine the intra-task information to guide a ranking distance calibration process.

Ranking in image retrieval. Image Retrieval (IR) is a classical computer vision task that aims to search in unlabelled gallery data in order to find those images that are most similar to a probe image. In IR, a ranking method [11,13,22,26] computes a rank list by a distance metric. Moreover, different re-ranking ideas [33, 44, 53] were proposed as a post-processing mechanism to improve the initial ranking result. For instance, a model [53] uses the concept of k-reciprocal nearest neighbours [31] to explore more hard positive samples. The enlarged k-nearest neighbours are then used to recompute a Jaccard distance as an auxiliary distance matrix. This idea was further adopted by others for IR [25,33,44].

In this work, we explore the concept of re-ranking in few-shot learning. Although related to the work of [47], our approach differs. The idea of [47] is to improve subgraph similarity by a graph view *solely in a pre-trained space*. In contrast, our idea is to calibrate a similarity distance *both* in a pre-trained space and a task-adaptive subspace. To optimise a feature representation, we use this calibrated distance to guide cross-domain knowledge transfer with a Kullback-Leibler divergence loss. The model of [47] does not learn cross-domain knowledge in re-ranking. It uses a MLP to meta-learn a subgragh refiner *in a single domain* with a cross entropy loss.

3. Methodology

Problem formulation. We start by defining a general FSL problem: given a source dataset $\mathcal{D}_s$ and a target dataset $\mathcal{D}_t$, where the classes in $\mathcal{D}_s$ and $\mathcal{D}_t$ are disjoint. FSL aims to address the C-way K-shot classification task $\mathcal{T}$ in $\mathcal{D}_t$ by leveraging the limited data in $\mathcal{T}$ and the prior knowledge learned from D_s containing lots of labelled images. In specific, a FSL task $\mathcal{T}$ contains a labelled support set $\mathcal{T}_s = \{\mathbf{I}_i, y_i\}_{i=1}^{C\times K}$ and an unlabelled query set $\mathcal{T}_q = \{\mathbf{I}_j\}_{j=1}^{C\times Q}$, where the images in $\mathcal{T}_s$ and $\mathcal{T}_q$ are both from the same C classes and K/Q denote the number of images per class in $\mathcal{T}_s$/$\mathcal{T}_q$. The goal of FSL is to recognise the unlabelled query set $\mathcal{T}_q$ when K is small. In the CD-FSL setup, $\mathcal{D}_s$ and $\mathcal{D}_t$ are from different domains, *e.g.* $\mathcal{D}_s$ contains many natural photos whilst the images in $\mathcal{D}_t$ are captured by remote sensing sensors.

Nearest prototype classifier. We firstly define the prototype classifier used in this work. Given a feature extractor f, we can extract the embedding $x = f(\mathbf{I})$ for image $\mathbf{I}$ in a FSL task $\mathcal{T}$. Nearest Prototype Classifier (NPC) first computes the prototypes $\hat{X} = \{\hat{x}_0, \cdots, \hat{x}_{C-1}\}$ for the C classes, where the prototype for class c is:

$$\hat{x}_c = \frac{1}{K}\sum_{i=1}^{K} x_i, \quad \text{where } x_i \in \mathcal{T}_s, y_i = c. \tag{1}$$

With the prototypes, the labels for x_j in $\mathcal{T}_q$ is assigned by:

$$y(x_j) = \text{argmin}_{\,c\in\{0,\cdots,C-1\}}\; d(x_j, \hat{x}_c), \tag{2}$$

where $d(\cdot)$ is a distance metric, *e.g.* Euclidean distance in this work, and $d(x_j, \hat{x}_c)$ is the distance between x_j and $\hat{x}_c$.

Overview. The key insight of this paper is to formulate the FSL as the Image Retrieval (IR) task by sharing the same angle with [40]. In [40], the authors propose to optimise the mean average precision for FSL. Furthermore, our view of "FSL as IR" also emphasises the importance of maximally leveraging all the available information in this low-data regime whilst concerns on the calibration of pairwise distances in FSL. In particular, this work follows this view for FSL and consider each sample in FSL as the probe data in IR and treat the whole FSL data as the gallery data. To this end, we propose a ranking distance calibration process for CD-FSL, and our key methodology is to repurpose the re-ranking to find the relevant images from the FSL task for a given image. We overview the proposed method in Fig. 2.

3.1. Ranking Distance Calibration (RDC)

Motivation. Previous works [31, 53] have suggested that discovering the k-reciprocal nearest neighbours within the gallery data can benefit the re-ranking result for image retrieval. This observation encourages us, when considering few-shot as image retrieval, to reuse this k-reciprocal nearest neighbour discovery process in [53] to calibrate the pairwise distances within FSL task. Fig. 1 gives an illustration: The k-reciprocal nearest neighbour discovery process finds more hard-positive samples for a given query sample. These hard-positive samples are then used to update each pairwise distance by k-reciprocal encoding and to further estimate a better distance by query expansion. These pairwise distances in a FSL task are calibrated and represented in a new distance matrix.

Let us now describe the re-ranking process in our ranking distance calibration. We further describe in details the

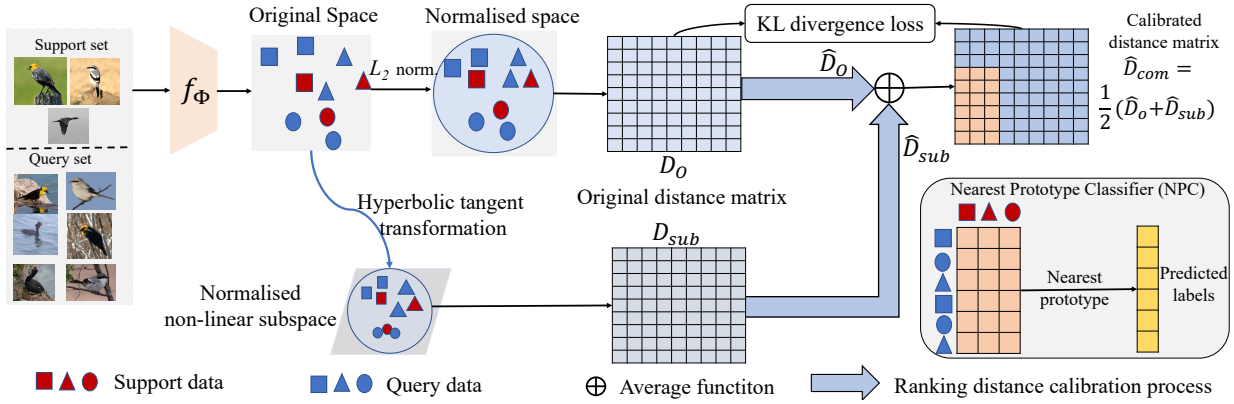


Figure 2. **An overview of the proposed ranking distance calibration pipeline.** f_Φ is the feature extractor pre-trained on the source dataset with a standard cross entropy loss. Our **RDC** method contains two parts: 1) a ranking process on the original space to calibrate the D_o to $\hat{D}_o$, and 2) a ranking process on a non-linear subspace to calibrate the D_{sub} to $\hat{D}_{sub}$. The proposed **RDC-FT** method uses a KL loss between D_o and $\hat{D}_{com}$ to fine-tune f_Φ. The NPC is used to classify the query data according to the pairwise distances, *i.e.* the calibrated distances $\hat{D}_{com}$ by RDC and the Euclidean distances on the embeddings fine-tuned with RDC-FT, respectively.

Jaccard distance computing process in the supplementary material. For a FSL task, we start by computing an original pairwise Euclidean distance matrix:

$$D_o = \begin{bmatrix} d_o(1,1) & \cdots & d_o(1,n) \\ \vdots & \ddots & \vdots \\ d_o(n,1) & \cdots & d_o(n,n) \end{bmatrix}, \text{ where } n = C\times(K+Q). \tag{3}$$

$d_o(i,j)$ is the Euclidean distance between $\|x_i\|_2$ and $\|x_j\|_2$, and $\|\cdot\|_2$ represents the l_2 normalisation. Referring to D_o, we can obtain the k-nearest neighbours set $R = [R_1(k), R_2(k), ..., R_n(k)]$, and $R_i(k)$ is the k-nearest neighbours of x_i. The re-ranking idea [53] is to expand the $R_i(k)$ by discovering more hard-positive samples for x_i. The expand process for $R_i(k)$ is guided by a k-reciprocal nearest neighbours algorithm [31] and the expanded ranking list $\hat{R}_i(k)$ is used to estimate a calibrated distance matrix.

k-reciprocal discovery and encoding. The principle of k-reciprocal nearest neighbour discovery is that if x_g is in $R_i(k)$, then x_i should also occur in $R_g(k)$ [31]. Given this assumption, we adopt the method of [53] for computing the expanded $\hat{R}_i(k)$, defined as:

$$\begin{aligned} &\hat{R}_i(k) \leftarrow R_i(k) \cup \mathcal{R}_g((1/2)k) \\ &s.t. \left|R_i(k) \cap R_g((1/2)k)\right| \geq (2/3)\left|R_g((1/2)k)\right|, \end{aligned} \tag{4}$$

where x_g is the sample in $\mathcal{R}_i(k)$ and $|\cdot|$ is the number of neighbours. In particular, to leverage effectively the labelled support data in FSL, we remove the labelled distractors from neighbours and further expand $\hat{R}_i(k)$ by uniting $\hat{R}_s(k)$, where x_s and x_i are both in the support set as well as from the same class. Finally, the expanded k-nearest neighbours set is $\hat{R} = [\hat{R}_1(k), \cdots, \hat{R}_n(k)]$. To assign larger weights to closer neighbours while smaller weights to farther ones, the $\hat{R}_i(k)$ is further used to encode the $\mathbf{d}_o(i,:) = [d_o(i,g_1), \ldots, d_o(i,g_n)]$ into a vector $\mathcal{V}_i = [\mathcal{V}_{i,g_1}, \mathcal{V}_{i,g_2}, ..., \mathcal{V}_{i,g_n}]$, where $\mathcal{V}_{i,g_i}$ is defined as the Gaussian kernel of the pairwise distance as:

$$\mathcal{V}_{i,g_q} = \begin{cases} e^{-d_o(x_i,x_{g_q})} & \text{if } g_q \in \hat{\mathcal{R}}_i(k) \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

After that, a query expansion strategy [4,31] is employed to integrate k_2 most-likely samples to update the feature of x_i by: $\mathcal{V}_i = (1/|\hat{R}_i(k_2)|)\sum_{g_q\in\hat{R}_i(k_2)} \mathcal{V}_{g_q}$, where $k_2 < k$.

Jaccard distance. Referring to [2, 53], the expanded ranked list $\hat{R}$ is used as contextual knowledge to compute a Jaccard distance matrix $D_J = \{d_J(i,g_q)|q \in [1,n]\}$ by:

$$d_J(i,g_q) = 1 - \frac{|\hat{R}_i(k) \cap \hat{R}_{g_q}(k)|}{|\hat{R}_i(k) \cup \hat{R}_{g_q}(k)|}. \tag{6}$$

Following the re-weighting method in [53], the number of candidates in the intersection and union set can be calculated as $|\hat{R}_{g_q}(k) \cap \hat{R}_i(k)| = ||\min(\mathcal{V}_i, \mathcal{V}_{g_q})||_1$ and $|\hat{R}_{g_q}(k) \cup \hat{\mathcal{R}}_i(k)| = ||\max(\mathcal{V}_i, \mathcal{V}_{g_q})||_1$, where min and max operate the element-based minimisation and maximisation for two input vectors, and $||\cdot||_1$ is l_1 norm. Then the Jaccard distance in Eq.(6) can be re-formulated as:

$$d_J(i,g_q) = 1 - \frac{\sum_{j=1}^{n} \min(\mathcal{V}_{i,g_j}, \mathcal{V}_{g_q,g_j})}{\sum_{j=1}^{n} \max(\mathcal{V}_{i,g_j}, \mathcal{V}_{g_q,g_j})}. \tag{7}$$

Distance calibration. The Jaccard distance exploits the contextual information to compute a relative distance in context within k-reciprocal neighbours. The original distance is an absolute distance in pre-trained Euclidean space. Therefore, combing D_J and D_o makes the distance more discriminative among neighbours by a weighting strategy:

$$\hat{D}_o = \lambda D_o + (1-\lambda) D_J, \tag{8}$$

where λ is a trade-off scalar to balance the two matrices.

3.2. RDC in Task-adaptive Non-linear Subspace

To further bridge the domain gap, we propose further improving the RDC in a non-linear subspace. We particular tailor a discriminative subspace to help calibrate the ranking in our CD-FSL task. The subspace is built upon the Principal Component Analysis (PCA) to extract crucial features from the original space. In specific, given the feature representations $X \in \Re^{n\times m}$ of a target FSL task, we have

$$X_{sub} = XP, \quad \text{where } \mathcal{P} \in \Re^{m\times p}, X_{sub} \in \Re^{n\times p}. \tag{9}$$

$\mathcal{P}$ is a transformation matrix mapping the feature with m dimensions to a reduced feature with p dimensions.

Hyperbolic tangent transformation. Generally, the PCA method can be directly used on the original embedding space. However, the original representation X is scattered due to the biased and less-discriminative embedding; thus the dimensional reduction easily causes the information loss problem. To remit this issue, we consider to transform the original representations to a compact and representative non-linear space. By using the idea of kernels, we use a hyperbolic tangent function to construct a task-adaptive non-linear subspace. Our non-linear PCA method first computes a feature-toward kernel function by:

$$\mathcal{K} = \tanh(X^T X), \ \mathcal{K} \in \Re^{m\times m}. \tag{10}$$

Then we use Singular Value Decomposition (SVD) to compute the eigenvalues U of $\mathcal{K}$ and select the most p-relevant eigenvalues U^p, formulating the transformation matrix $\mathcal{P} = U^p$. To this end, a task-adaptive non-linear subspace X_{sub} is construct by Eq.(9) and Eq.(10).

Complementary distance calibration. Our distance calibration process is space-agnostic and can be applied in the original linear space (in Sec. 3.1) and a non-linear subspace (in Sec. 3.2). The original space has higher dimensions consist of full information but also disturbed by noisy task-irrelevant features, while the non-linear subspace reduce some task-irrelevant signal but loss some information. Our RDC method co-leverages the calibrated distances in the two spaces to capture a robust and complementary distance matrix $\hat{D}_{com} = 0.5(\hat{D}_o + \hat{D}_{sub})$. How to compute RDC in two spaces is described in lines 4-8 of Alg. 1.

Remark. As the hyperbolic non-linear space has larger capability than Euclidean space [7, 14, 48], it can alleviate the information loss caused by the dimensionality reduction. Hence, we use a hyperbolic tangent transformation to map the source domain linear space to a non-linear space. We note that the subspace learning has been preliminary explored in FSL work [35,50] to learn task-adaptive or class-adaptive subspace. Critically, our subspace construction method differs from [35, 50] as our method does not need the sophisticated episode training process.

Algorithm 1: Ranking Distance Calibration with Fine-Tuning (RDC-FT)

Data: pre-trained feature extractor f_Φ; support set $\mathcal{T}_s$; query set $\mathcal{T}_q$; RDC: k_1, k_2, λ; Fine-tune: epochs T, learning rate β, τ, α.
Result: Fine-tuned feature extractor $f_{\hat{\Phi}}$.

```
/* RDC-FT: optimise f_Φ by RDC */
```
1 Initialise $\hat{\Phi} = \Phi$;
2 **for** t *in* T **do**
3 Extract embeddings for $\mathcal{T}_s$ and $\mathcal{T}_q$ by $f_{\hat{\Phi}}$;
```
/* RDC: distance calibration */
```
4 Compute original distances D_o by Eq.(3) ;
5 Compute calibrated distances $\hat{D}_o$ by Eq.(8) ;
6 Construct a non-linear subspace by Eq.(9, 10) ;
7 Calibrate distance in subspace $\hat{D}_{sub}$ by Eq.(8) ;
8 Compute $\hat{D}_{com} = 0.5 * (\hat{D}_o + \hat{D}_{sub})$;
```
/* FT: optimise f_Φ̂ with D̂_com */
```
9 Get the softened distribution $\mathbf{p}^{D_o}(\tau)$/$\mathbf{p}^{\hat{D}_{com}}(\tau)$;
10 Compute KL loss by Eq.(12);
11 $\hat{\Phi} \leftarrow \hat{\Phi} - \beta \bigtriangledown_{\hat{\Phi}} L_{KL}(\mathbf{p}^{D_o}(\tau), \mathbf{p}^{\hat{D}_{com}}(\tau))$;
12 **end**

3.3. Fine-tuning with RDC

As $\hat{D}_{com}$ provides a more robust and discriminative distance matrix, it is natural to ask whether this type of calibration knowledge can be used to optimise the feature extractor. To achieve this, we fine-tune the feature extractor by iteratively mapping the original distance distribution to the calibrated distance distribution, formulating a RDC with Fine-Tuning (RDC-FT) method as in Alg. 1.

Expanded k-reciprocal list as attention. As in Eq. (6), the expanded ranking list $\hat{R}(k)$ is used to re-compute the pairwise distances. The calibrated pairwise distances in $\hat{R}(k)$ are more robust than these not in $\hat{R}(k)$. Thus the $\hat{R}(k)$ can naturally be used as an attention mask $\mathcal{M}$. In particular, a $\mathcal{M}$ is computed by

$$\mathcal{M}_i^{g_q} = \begin{cases} 1+\alpha & \text{if } g_q \in \hat{\mathcal{R}}_i(k) \\ 1 & \text{otherwise,} \end{cases} \tag{11}$$

where α is an attention scalar. During the fine-tuning process, the $\mathcal{M}$ is used to re-weight the distance matrices D_o and $\hat{D}_{com}$ as $\mathcal{M} \cdot D_o$ and $\mathcal{M} \cdot \hat{D}_{com}$, respectively.

Choices of loss functions. To achieve the distance distribution alignment, Mean Squared Error (MSE) loss and Kullback-Leibler (KL) divergence loss are alternatives. The MSE loss prefers to directly learn towards the target distance while KL divergence loss focuses on the distribution matching [15]. As KL loss learns this mapping process in a softening way, it is a better way to embed the calibration

Method	5-way 1-shot								
	CUB	Cars	Places	Plantae	CropDisease	EuroSAT	ISIC	ChestX	Ave.
ProtoNet [36]	38.66±0.4	31.34±0.3	47.89±0.5	31.75±0.4	51.22±0.5	52.93±0.5	29.20±0.3	21.57±0.2	38.07
NPC	38.47±0.4	33.27±0.3	40.84±0.4	36.77±0.4	64.76±0.5	51.45±0.5	29.46±0.3	22.74±0.2	39.72
NPC+l_2 norm	43.67±0.4	35.76±0.4	48.67±0.4	39.15±0.5	66.62±0.5	60.85±0.5	31.52±0.3	**22.87±0.2**	43.64
RDC (ours)	**48.68±0.5**	**38.26±0.5**	**59.53±0.5**	**42.29±0.5**	**79.72±0.5**	**65.58±0.5**	**32.33±0.3**	22.77±0.2	**48.65**
Method	5-way 5-shot								
	CUB	Cars	Places	Plantae	CropDisease	EuroSAT	ISIC	ChestX	Ave.
ProtoNet [36]	57.55±0.4	43.98±0.4	68.05±0.4	46.18±0.4	79.98±0.3	75.36±0.4	39.98±0.3	24.19±0.2	54.41
NPC	60.48±0.4	51.16±0.4	67.74±0.4	53.34±0.4	86.95±0.3	74.27±0.4	39.26±0.3	26.17±0.2	57.42
NPC+l_2 norm	63.23±0.4	51.92±0.4	69.95±0.4	55.76±0.4	87.76±0.3	76.29±0.4	41.08±0.3	**26.31±0.2**	59.03
RDC (ours)	**64.36±0.4**	**52.15±0.4**	**73.24±0.4**	**57.50±0.4**	**88.90±0.3**	**77.15±0.4**	**41.28±0.3**	25.91±0.2	**60.06**

Table 1. **Comparisons with the baselines using NPC classifier *w/o* fine-tuning.** The classification accuracies on 8 datasets with **ResNet10** as the backbone. RDC exploits the labelled support data and the unlabelled query data in a FSL task. **Bold:** The best scores.

knowledge into the representations. Thus we use KL loss:

$$\mathcal{L}_{KL}(\mathbf{p}^{D_o}(\tau), \mathbf{p}^{\hat{D}_{com}}(\tau)) = \tau^2 \sum_j \mathbf{p}_j^{\hat{D}_{com}}(\tau) \log \frac{\mathbf{p}_j^{\hat{D}_{com}}(\tau)}{\mathbf{p}_j^{D_o}(\tau)}, \tag{12}$$

where τ is the temperature-scaling hyper-parameter, $\mathbf{p}_j^{D_o}(\tau)$ and $\mathbf{p}_j^{\hat{D}_{com}}(\tau)$ are the softened distributions of the re-weighted distances matrices $\mathcal{M} \cdot D_o$ and $\mathcal{M} \cdot \hat{D}_{com}$. Given a vector $\mathbf{d}(i,:)$ in the distance matrix D, the softened distribution $\mathbf{p}^{d(i,:)}(\tau)$ is denoted by $\mathbf{p}^{d(i,l)}(\tau) = \frac{\exp(d(i,l)/\tau)}{\sum_{j=1}^{n} \exp(d(i,j)/\tau)}$, where $d(i,l)$ is the l-th value of $\mathbf{d}(i,:)$.

4. Experiments

Dataset. Following the benchmarks in [20, 45], we used *miniImageNet* as the source domain and another eight datasets , *i.e. CUB*, *Cars*, *Places Plantae*, *CropDisease*, *EuroSAT*, *ISIC* and *ChestX*, as target domains. In specific, *miniImageNet* [42] is a subset of of *ILSVRC-2012*. *CUB*, *Cars*, *Places* and *Plantae* are the target domains proposed in [41] for the evaluation on natural image domains, while *CropDisease*, *EuroSAT*, *ISIC* and *ChestX* are four domains proposed in [10] for generalising the model to domains with different visual characteristics. For all experiments, we resized all the images to 224×224 pixels and used data augmentations in [41, 45] as image transformation.

Evaluation protocol. We followed the evaluation protocols in [45] to evaluate our method on CD-FSL. In specific, for each target domain, we randomly selected 2000 FSL tasks and each task contains 5 different classes. Each class has 1/5 support labelled data and additional 15 unlabelled data for evaluation the performance, formulating the 5-way 1/5-shot CD-FSL problem. In all experiments, we reported the mean classification accuracy as well as 95% confidence interval on the query set of each domain. For comprehensive comparison, we listed the average accuracy (shown as Ave. in Tab. 1 2 and 5) of 8 domains.

Implementation details. Following previous works [10, 41, 45], we used a ResNet10 as feature extractor. Further, we used the same hyper-parameters for the experiments on different domains to fairly validate the generalisation ability. In specific, the feature extractor are pre-trained for 400 epochs on the base classes of *mini*ImageNet with an Adam optimiser. We set the learning rate as 0.001 and the batch size as 64. For our RDC method, we set $k = 10$, $k_2 = 8$ and $\lambda = 0.5$, and the reduced dimension p was set as 64 for the non-linear subspace. For the fine-tuning stage in RDC-FT, we set the attention scalar $\alpha = 0.5$, temperature $\tau = 3$ and $T = 20$ epochs for model training using an Adam optimiser with learning rate β as 0.001.

4.1. Comparison with baselines

As our methods are based on a simple NPC classifier, here we start by comparing our RDC method with some baseline methods which also use a NPC classifier and do not need fine-tuning on a target domain. These baselines are: NPC that uses a NPC classifier on the pre-trained embedding, NPC+l_2 norm which utilises a NPC classifier on a l_2 normed feature embeddings, and ProtoNet [36] that meta-learns a task-agnostic NPC classifier on *mini*ImageNet. The results in Tab. 1 show that RDC largely outperforms these baselines, boosting the simple NPC classifier to a strong one. In particular, the performance on 1-shot learning is improved notably with $5 \sim 10\%$ increases on the *Ave.* accuracy compared to the baselines. This observation indicates that RDC is efficient to calibrate the distances by fully-leveraging the task information, *i.e.* the labelled support data and unlabelled query data. We also note that the improvement on 5-shot is not as large as that on 1-shot. The reason is that the prototypes for the NPC classifier is more robust under many-shot setting, thus the original distances are less-biased and this calibration process improves less when the embedding is fixed. This limitation can be remitted by using the fine-tuning stage of our RDC-FT method.

4.2. Comparison with state-of-the-art methods

We further compared our RDC-FT method with State-of-The-Art (SoTA) methods: 1) meta-learners: GNN-FT [41] that meta-trains a GNN [34] model with an additional Feature Transformation layer, GNN-LRP which uses a Layer-

Method	5-way 1-shot								
	CUB	Cars	Places	Plantae	CropDisease	EuroSAT	ISIC	ChestX	Ave.
GNN+FT† [41]	45.50±0.5	32.25±0.4	53.44±0.5	32.56±0.4	60.74±0.5	55.53±0.5	30.22±0.3	22.00±0.2	41.53
GNN+LRP† [37]	43.89±0.5	31.46±0.4	52.28±0.5	33.20±0.4	59.23±0.5	54.99±0.5	30.94±0.3	22.11±0.2	41.01
TPN+ATA* [45]	50.26±0.5	34.18±0.4	57.03±0.5	39.83±0.4	77.82±0.5	65.94±0.5	34.70±0.4	21.67±0.2	47.68
Fine-tuning† [10]	43.53±0.4	35.12±0.4	50.57±0.4	38.77±0.4	73.43±0.5	66.17±0.5	34.60±0.3	22.13±0.2	45.54
ConFT‡ [6]	45.57±0.8	**39.11±0.7**	49.97±0.8	43.09±0.8	69.71±0.9[1]	64.79±0.8[1]	34.47±0.6[1]	**23.31±0.4**[1]	46.25
RDC-FT*‡ (ours)	**51.20±0.5**	**39.13±0.5**	**61.50±0.6**	**44.33±0.6**	**86.33±0.5**	**71.57±0.5**	**35.84±0.4**	22.27±0.2	**51.53**
Method	5-way 5-shot								
	CUB	Cars	Places	Plantae	CropDisease	EuroSAT	ISIC	ChestX	Ave.
GNN+FT† [41]	64.97±0.5	46.19±0.4	70.70±0.5	49.66±0.4	87.07±0.4	78.02±0.4	40.87±0.4	24.28±0.2	57.72
GNN+LRP† [37]	62.86±0.5	46.07±0.4	71.38±0.5	50.31±0.4	86.15±0.4	77.14±0.4	44.14±0.4	24.53±0.3	57.82
TPN+ATA* [45]	65.31±0.4	46.95±0.4	72.12±0.4	55.08±0.4	88.15±0.5	79.47±0.3	45.83±0.3	23.60±0.2	59.57
Fine-tuning†‡ [10]	63.76±0.4	51.21±0.4	70.68±0.4	56.45±0.4	89.84±0.3	81.59±0.3	49.51±0.3	25.37±0.2	61.06
ConFT‡ [6]	**70.53±0.7**	**61.53±0.7**	72.09±0.7	**62.54±0.7**	90.90±0.6[1]	81.52±0.6[1]	50.79±0.6[1]	**27.50±0.5**[1]	**64.68**
NSAE(CE+CE)‡ [20]	68.51±0.8	54.91±0.7	71.02±0.7	59.55±0.8	93.14±0.5	83.96±0.6	**54.05±0.6**	27.10±0.4	64.03
RDC-FT*‡ (ours)	67.77±0.4	53.75±0.5	**74.65±0.4**	60.63±0.4	**93.55±0.3**	**84.67±0.3**	49.06±0.3	25.48±0.2	63.70

Table 2. **Comparisons with SoTA methods**. The 5-way 1/5-shot classification accuracies on 8 domains with **ResNet10** as the backbone. † indicates the result reported in [45]. ‡ denotes fine-tuning in the target domain. * denotes using both the labelled support data and the unlabelled query data. $(\cdot)^1$ denotes our reproduced results using the official released code from [6]. **Bold** denotes the best scores.

wise Relevance Propagation to guide the GNN model training, and TPN+ATA [45] that meta-learns TPN [24] with Adversarial Task Augmentation; 2) fine-tuning methods: a general Fine-tuning [10] method, ConFT [6] that fine-tunes model reusing the base classes, and NSAE [20] which pre-trains and fine-tunes model with an additional autoencoder task to improve the model generalisation. From Tab. 4.2, we observe that our RDC-FT method is superior to the SoTA methods on the 1-shot learning and competitive to SoTAs on the 5-shot learning. Also, we notice that the performance is not superior to ConFT and NSAE methods for the 5-shot learning. The behind reasons are: 1) our method explores the task information in an unsupervised way while the others focus on fine-tuning with more labelled data; thus these methods benefit a lot from the 5-shot setting. 2) ConFT reuses more data from base classes for model fine-tuning. Thus the similar classes between source and target domain, *e.g.* birds, cars, help to build more robust decision boundaries when model learning on related target domains, *e.g. CUB*, *Cars*. But this approach requires more data and expensive computing resources. 3) NASE adopted an autoencoder to implicitly augment data to pre-train a generalisable model, and our method is theoretically orthogonal to this method for solving the CD-FSL problem.

4.3. Ablation study

Component analysis. To investigate the efficacy of components in RDC-FT, we ablated the contribution of each element in RDC-FT: RDC *w/o* subspace, RDC (in two spaces), RDC-FT *w/o* subspace and RDC-FT (in two spaces). As in Tab. 3, a simple RDC process without subspace learning, which calibrates the distances only on the pre-trained representation, largely boosts the baseline NPC classifier by 7.99% (1.55%) improvement on 1-shot (5-shot). The fine-tuning process, as in results of RDC-FT *w/o* subspace, can increase the improvement by an iterative mapping process,

Method	5-way 1-shot	5-way 5-shot
Baseline NPC	39.72	57.42
+RDC *w/o* subspace	47.71(↑7.99%)	58.97(↑1.55%)
+RDC	48.65(↑8.93%)	60.06(↑2.64%)
+RDC-FT *w/o* subspace	51.03(↑11.31%)	62.91(↑5.49%)
+RDC-FT	51.53(↑11.81%)	63.70(↑6.28%)

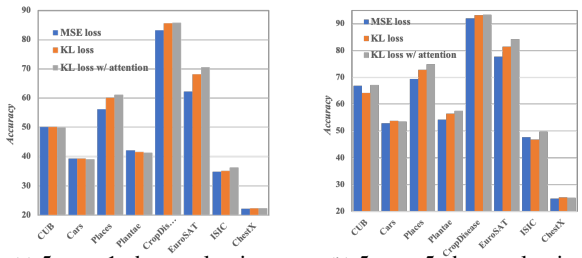
Table 3. **Component analysis of the proposed RDC-FT method.** The results are the average accuracies of 8 target domains. (↑ γ %) indicates γ% improvement compared to the NPC baseline.

Method	5-way 1-shot					
	N/A	linear	Gaussian	Poly.	Sigmoid	Ours
NPC+l_2 norm	43.72	45.54	45.53	45.49	45.59	**45.81**
RDC	47.77	47.41	47.38	47.32	47.46	**48.36**
Method	5-way 5-shot					
	N/A	linear	Gaussian	Poly.	Sigmoid	Ours
NPC+l_2 norm	59.03	58.50	58.52	58.43	58.56	**59.72**
RDC	59.09	58.83	58.82	58.75	58.86	**59.83**

Table 4. **Comparisons of different subspaces.** The average accuracies of 8 target domains by using NPC+l_2 norm and RDC methods on the subspaces constructed by KPCA with different kernels. N/A represents the original representation without subspace.

achieving 11.31% (5.49%) improvement on 1-shot (5-shot). Interestingly, we observe that the contribution of subspace for RDC (↑0.94%) is larger than that for RDC-FT (↑0.50%) on 1-shot. This indicates that fine-tuning process can gradually alleviate the bias of pre-trained representations, thus the benefit of subspace becomes less in RDC-FT.

Comparison of different PCA methods. We compared our non-linear subspace to the subspaces constructed by Kernel PCA (KPCA) methods with different kernel types (linear, Gaussian, Polynomial and Sigmoid). For fair comparison, the dimensions of subspaces are set as 64 and we used the default parameters of KPCA methods in *scikit-learn* [29]. Table 4 shows that our non-linear subspace performs better than others. In particular, without RDC method, the KPCA methods can largely improve the performance (compared to N/A) on 1-shot learning, but the results of KPCA methods are just competitive to the original space when applying RDC on different subspaces. How-



(a) 5-way 1-shot evaluation. (b) 5-way 5-shot evaluation.

Figure 3. **Results of RDC-FT with different loss functions**. The evaluation is conducted on 2000 tasks from 8 target domains.

ever, our non-linear subspace achieves consistent and stable improvement both with and *w/o* RDC method, verifying the robustness and superiority of our non-linear subspace.

Effect of loss choices. We evaluated the performance of RDC-FT with different loss functions. The results in Fig. 3 show that these three losses achieve competitive performance on *CUB*, *Cars*, *ISIC* and *chestX*, while the KL loss performs mostly better than MSE loss on *Places*, *Plantae*, *EuroSAT*. These observations suggest the superiority of mapping the distance matrix in softened distributions. We conjecture this should attribute to the softening process which can alleviate the negative effect of the calibrated distances. Moreover, we note that the performance of KL loss can be further improved by an attention strategy on the distance matrices, verifying the efficacy of employing the expanded k-nearest neighbours list as an attention reference.

Visualisation. To intuitively understand the advantages and limitations of the proposed RDC and RDC-FT methods, we used t-SNE [27] to display the classification results of three different methods – a NPC classifier, RDC, and RDC-FT, on a 5-way 1-shot FSL task from *CropDisease*. The results in Fig. 4 show: 1) RDC can correct some misclassified samples that are near to the support exemplars, *i.e.* the samples in red solid rectangles in plots(II)&(III). However, RDC cannot address effectively the misclassified samples between different support exemplars, *i.e.* the failure cases in plot(III). 2) From the samples in red dashed rectangles of plots(I)&(IV), it is evident that RDC-FT can calibrate the distance-based distributions in the representational space, encouraging the feature representations to have smaller intra-class variations and greater inter-class margins, resulting in fine-tuned representations being more discriminative for classification. 3) The failure cases of RDC, *i.e.* M-R1, M-R3, M-R4, and M-R5 in plot(III), can be correctly classified by RDC-FT with a simple NPC classifier, as shown in plot(V). This verifies the superiority of RDC-FT that gradually embeds the calibration information to the representational space.

Incorporate with other method. As RDC is a post-processing method, it can flexibly combine with other methods. Here we employed RDC on a general data augmentation method [49]. The results in Tab. 5 indicate that

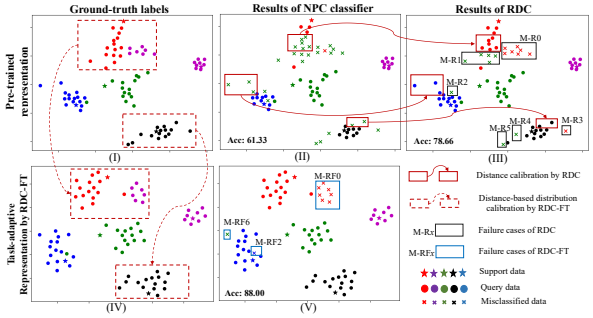


Figure 4. **T-SNE visualisation of 5-way 1-shot task from *CropDisease*.** We show the classification results with the NPC classifier and our RDC method in different representations, *i.e.* Pre-trained representations and task-adaptive representation by RDC-FT. The different colours of the points (round/cross/star) represent the ground-truth labels or the labels assigned by NPC/RDC.

Method	5-way 1-shot							
	CUB	Cars	Places	Plantae	Crop.	Euro.	ISIC	ChestX
NPC+l_2+DA†	40.49	34.53	45.86	36.78	71.94	64.29	32.91	**22.02**
RDC+DA†	**45.98**	**38.17**	**58.30**	**40.73**	**84.73**	**71.59**	**34.44**	22.03
Method	5-way 1-shot							
	CUB	Cars	Places	Plantae	Crop.	Euro.	ISIC	ChestX
NPC+l_2+DA†	65.23	53.61	68.91	57.72	91.66	83.09	50.63	**26.03**
RDC+DA†	**67.68**	**55.13**	**73.06**	**60.34**	**93.41**	**84.67**	**51.53**	25.79

Table 5. **Incorporate RDC with others.** † denotes using NPC *w/* l_2 norm on the model fine-tuned with data augmentation in [49].

RDC can achieve consistent improvement on other method, showing its generalisable ability. Currently we cannot evaluate our method on [20, 21] until their code is released.

5. Conclusions

In this paper, we proposed a Ranking Distance Calibration (RDC) method to calibrate the biased distances in CD-FSL. The calibration process is achieved by a re-ranking method with a k-reciprocal discovery and encoding process. As the pre-trained linear embedding is biased for target domain, we further proposed a non-linear subspace followed by a calibration process on it. Our RDC method averages the calibrated distances on the two spaces to a robust distance matrix. Moreover, we introduced a RDC-FT method to fine-tune the embedding with the calibrated distances, yielding a discriminative representation for CD-FSL task.

Limitation. As the image retrieval perspective of our approach is to discover the task information unsupervised, exploring comprehensive leveraging of the label information and the task information should be considered, especially in the many-shot cases, *e.g.* 5-shot.

Ackowledgements. This work was supported by Vision Semantics Limited, the Alan Turing Institute Fellowship, the China Scholarship Council, NSFC under Grant (No. 62076067), and SMSTM Projects (2018SHZDZX01 and 2021SHZDZX0103). We shall also thank Jun Liu and Bin-Bin Gao of Tencent YouTu Lab for helpful discussions and GPU support.

References

[1] Thomas Adler, Johannes Brandstetter, Michael Widrich, Andreas Mayr, David Kreil, Michael Kopp, Günter Klambauer, and Sepp Hochreiter. Cross-domain few-shot learning by representation fusion. *arXiv preprint arXiv:2010.06498*, 2020. 3

[2] Song Bai and Xiang Bai. Sparse contextual activation for efficient visual re-ranking. *IEEE TIP*, 25(3):1056–1069, 2016. 4

[3] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 2

[4] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, pages 1–8. IEEE, 2007. 4

[5] Debasmit Das, Sungrack Yun, and Fatih Porikli. ConfeSS: A framework for single source cross-domain few-shot learning. In *ICLR*, 2022. 3

[6] Rajshekhar Das, Yu-Xiong Wang, and Jose MF Moura. On the importance of distractors for few-shot classification. In *ICCV*, pages 9030–9040, 2021. 3, 7

[7] Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Kernel methods in hyperbolic spaces. In *ICCV*, pages 10665–10674, 2021. 5

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 1, 2

[9] Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. *ACM MM*, pages 5326–5334, 2021. 1, 3

[10] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *ECCV*, pages 124–141. Springer, 2020. 1, 2, 6, 7

[11] Jingrui He, Mingjing Li, Hong-Jiang Zhang, Hanghang Tong, and Changshui Zhang. Manifold-ranking based image retrieval. In *ACM MM*, pages 9–16, 2004. 3

[12] Zhengdong Hu, Yifan Sun, and Yi Yang. Switch to generalize: Domain-switch learning for cross-domain few-shot classification. In *International Conference on Learning Representations*, 2022. 1

[13] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, pages 1062–1070, 2015. 3

[14] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, pages 6418–6428, 2020. 5

[15] Taehyeon Kim, Jaehoon Oh, NakYil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. *IJCAI*, 2021. 5

[16] Jianyi Li and Guizhong Liu. Few-shot image classification via contrastive self-supervised learning. *arXiv preprint arXiv:2008.09942*, 2020. 2

[17] Pan Li, Yanwei Fu, and Shaogang Gong. Regularising knowledge transfer by meta functional learning. In *IJCAI*, 2021. 2

[18] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *ICCV*, pages 8886–8895, 2021. 1

[19] Pan Li, Guile Wu, Shaogang Gong, and Xu Lan. Semi-supervised few-shot learning with pseudo label refinement. In *ICME*, pages 1–6. IEEE, 2021. 2

[20] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *ICCV*, pages 9424–9434, October 2021. 1, 2, 3, 6, 7, 8

[21] Bingyu Liu, Zhen Zhao, Zhenpeng Li, Jianan Jiang, Yuhong Guo, and Jieping Ye. Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. *arXiv preprint arXiv:2005.08463*, 2020. 2, 8

[22] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *CVPR*, pages 441–448, 2013. 3

[23] Lu Liu, William L Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. In *ICLR*, 2021. 1

[24] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sungju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *ICLR*, 2019. 7

[25] Yong Liu, Lin Shang, and Andy Song. Adaptive re-ranking of deep feature for person re-identification. *arXiv preprint arXiv:1811.08561*, 2018. 3

[26] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *ICIP*, pages 3567–3571. IEEE, 2013. 3

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. 9(Nov):2579–2605, 2008. 8

[28] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. Boil: Towards representation change for few-shot learning. In *ICLR*, 2021. 2

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 7

[30] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *ICLR*, 2021. 3

[31] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR*, pages 777–784. IEEE, 2011. 3, 4

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

[33] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, pages 420–429, 2018. 3

[34] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *ICLR*, 2019. 1, 6

[35] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *CVPR*, pages 4136–4145, 2020. 5

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017. 1, 2, 6

[37] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. Explanation-guided training for cross-domain few-shot classification. In *ICIP*, pages 7609–7616. IEEE, 2021. 7

[38] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018. 1, 2

[39] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, pages 266–282. Springer, 2020. 2

[40] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *NeurIPS*, pages 2252–2262, 2017. 3

[41] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020. 1, 2, 3, 6, 7

[42] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *NeurIPS*, 29:3630–3638, 2016. 6

[43] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *NeurIPS*, 2018. 1

[44] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*, pages 6449–6458, 2020. 3

[45] Haoqing Wang and Zhi-Hong Deng. Cross-domain few-shot classification via adversarial task augmentation. *IJCAI*, 2021. 1, 3, 6, 7

[46] Guile Wu and Shaogang Gong. Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In *ICCV*, pages 6484–6493, October 2021. 1

[47] Shell Xu Hu Othman Sbai Mathieu Aubry Xi Shen, Yang Xiao. Re-ranking for image retrieval and transductive few-shot classification. *NeurIPS*, 2021. 3

[48] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *CVPR*, pages 12465–12474, 2021. 5

[49] Jia-Fong Yeh, Hsin-Ying Lee, Bing-Chen Tsai, Yi-Rong Chen, Ping-Chia Huang, and Winston H Hsu. Large margin mechanism and pseudo query set on cross-domain few-shot learning. *arXiv preprint arXiv:2005.09218*, 2020. 8

[50] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *ICML*, pages 7115–7123. PMLR, 2019. 5

[51] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, pages 12203–12213, 2020. 2

[52] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *NeurIPS*, volume 33, pages 14435–14447, 2020. 1

[53] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017. 3, 4

[54] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 1