Class-wise Metric Scaling for Improved Few-Shot Classification

Ge Liu, Linglan Zhao, Wei Li, Dashan Guo and Xiangzhong Fang
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
{liu.ge, llzhao, liweihfyz, dmlab_gds, xzfang}@sjtu.edu.cn

Abstract

Few-shot classification aims to generalize basic knowledge to recognize novel categories from a few samples. Recent centroid-based methods achieve promising classification performance with the nearest neighbor rule. However, we consider that those methods intrinsically ignore per-class distribution, as the decision boundaries are biased due to the diversity of intra-class variances. Hence, we propose a class-wise metric scaling (CMS) mechanism, which can be applied to both training and testing stages. Concretely, metric scalars are set as learnable parameters in the training stage, helping to learn a more discriminative and transferable feature representation. As for testing, we construct a convex optimization problem to generate an optimal scalar vector for refining the nearest neighbor decisions. Besides, we also involve a low-ranking bilinear pooling layer for improved representation capacity, which further provides significant performance gains. Extensive experiments are conducted on a series of feature extractor backbones, datasets, and testing modes, which have shown consistent improvements compared to prior SOTA methods, e.g.*, we achieve accuracies of* $66.64\ \%$ *and* $83.63\ \%$ *for 5-way 1-shot and 5-shot settings on the* mini-*ImageNet, respectively. Under the semi-supervised inductive mode, results are further up to* $78.34\ \%$ *and* $87.53\ \%$*, respectively.*

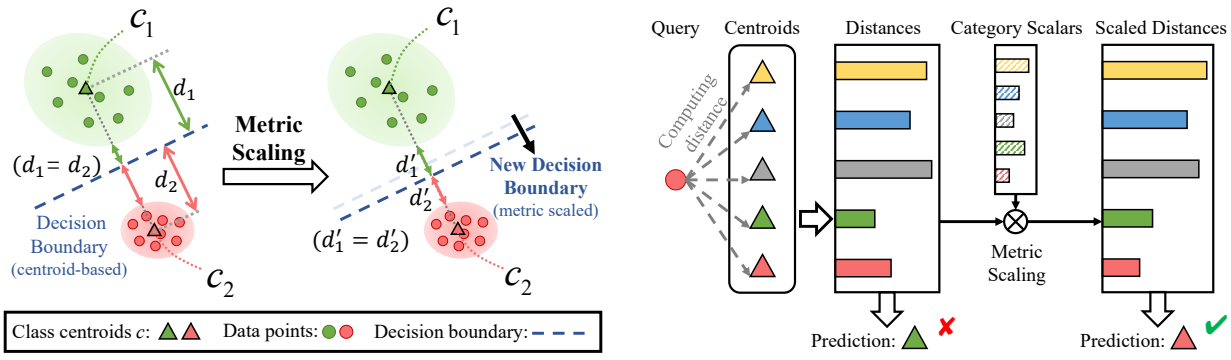
1. Introduction

With the availability of enormous labeled data, deep learning has achieved remarkable performance on multiple computer vision tasks, such as object detection [13], recognition [14], and segmentation [27]. However, manually collecting and labeling massive amounts of data is very expensive and time-consuming. In contrast, humans need only a few examples to learn a new concept, while this still remains a challenge for machines. As a result, few-shot learning has attracted widespread attention in the recent machine learning community.

Few-shot classification aims to generalize basic knowledge learned from a large-scale available base set to recognize new classes with only a few labeled samples. Recently nearest-centroid based methods [32, 39, 6, 7, 4] have attracted considerable attention due to their simplicity and effectiveness. The essential proceeding of those methods is to pre-train a feature extractor, where training manners can be meta-learning [39, 41] or just conventional supervised training with a learnable classification layer [32, 6, 7]. For addressing a target few-shot task, a nearest-centroid classifier is built with the feature embeddings of the few annotated samples. Generally, each class centroid can be calculated by the mean of the embeddings that belong to the class. It is noteworthy that the nearest-centroid classifier only stores a single centroid vector to represent a category. Such limited representations become an advantage in the low-shot regime, as it effectively alleviates overfitting.

Although remarkable performance has been achieved, we argue that centroid-based methods ignore the distribution difference among classes, where classification results only based on point-to-point distances are biased. Our motivation can be described by a geometric interpretation of a binary-category classification scenario in Fig.1(a). Under the centroid-based metric with the nearest neighbor rule, the decision boundary (the dashed line) lies in the middle of two class centroids. However, the cluster of "Class 1" is more widely distributed and has larger intra-variance than "Class 2". Intuitively, a more reasonable decision boundary should lean to "Class 2". In order to refine the biased classification decision, we propose a simple yet effective *Class-wise Metric Scaling* (CMS) mechanism, which multiplies the vanilla class-agnostic distances with class-wise metric scalars w.r.t. the class distributions, as can be seen in Fig. 1(b). So the diversity of intra-class variances can be implicitly considered to improve the classification accuracy.

The CMS is beyond a trivial modification. It bears several essential advantages comparing with prior few-shot approaches. First, involving the per-class scaling according to class distribution information is simple and straightforward for any regular distance metric, so there is no need to change the existing training pipelines or distance functions. Second, it can be mathematically proved that the scalar in CMS is approximately equivalent to the reciprocal of the correspond-



(a) A binary-category classification scenario

(b) Class-wise Metric Scaling mechanism

Figure 1. Interpretation of class-wise metric scaling mechanism. Best viewed in color. (a) The decision boundary is the median line of two class centroids for the centroid-based metric with the Nearest Neighbor rule. After considering class distributions, a reasonable decision boundary approximately leans to the middle of the margin of two class distributions. (b) For a 5-way classification task, 5 scalars corresponding to 5 classes are used for refining the nearest distance metric.

ing class variance. In this case, the classifier is regarded as a special formulation of the *Gaussian Mixture Model* in Section 3.4. So extending the non-scaled (or equal-scaled) nearest-centroid classifier to the CMS-based one actually advocates more flexible class representations. Moreover, we apply CMS to both meta-training and meta-testing stages, while most prior scaling approaches only benefit the embedding learning during meta-training. Specifically, the positive effects of CMS can be attributed to not only the better feature representation during the meta-training phase but also a task-specific adaptation process in meta-testing. Besides, in order to further improve feature representation ability, we also incorporate a low-rank bilinear pooling layer to the end of the feature extractor. Experimental results also show favorable improvements with considerable margins.

This work makes main contributions as follow:

- We propose a class-wise metric scaling mechanism to address the few-shot classification by improving the NCC classifier. Consistent improvements and state-of-the-art results have been achieved across multiple feature extractor backbones and datasets, *e.g.*, on the *mini*-ImageNet we achieve accuracies of 66.64 % and 83.63 % for 5-way 1-shot and 5-shot settings, respectively.
- In addition to improved generic few-shot classification, CMS shows more significant superiority in dealing with the cross-domain scenario, *e.g.*, our performance is better than the prior SOTA method [1] by 9.8% and 8.2% for 5-way 1-shot and 5-shot settings, respectively.
- We extend our approach to the other two types of semi-supervised few-shot learning, i.e., semi-supervised transductive (SS-T) and inductive (SS-I) modes. With the unlabeled auxiliary data to enrich the class distributions, significant performance gains can be observed, *e.g.*, on 1-shot/5-shot SS-I mode CMS obtains accuracy of 78.34%/87.53% which improves 11.7%/3.9% compared to the standard few-shot mode.

2. Related works

Meta-learning. Few-shot learning has already made notable progress by the appearance of meta-learning with the episodic training mechanism [41]. Most typical episodic-trained methods are based on model optimization under the spirit of fast adaptation to new tasks and alleviating overfitting [8, 30, 35]. However, those methods are time-consuming due to model fine-tuning in the testing stage.

Metric-based few-shot Learning. Our approach is more related to metric-based methods which aim to learn a deep transferable representation first, and then generalize it to recognize novel classes. We can further divide these methods into two categories according to different training procedures. One type learns particular distance metrics based on the episodic training, including Matching Networks [41], Prototypical Networks [39], Relation Networks [40], TADAM [31], Covariance Metric Networks [20], DeepEMD [49]. The other type [11, 12, 22, 32, 6, 7] pre-trains a feature extractor combined with a classification layer in the standard supervised learning manner, and then builds nearest-centroid classifiers to address target few-shot tasks. In particular, some previous methods additionally apply task adaptation modules [11, 12, 22] for obtaining preciser class centroids, while other works [4, 6] show that it is also proper to build mean-centroid classifiers with a well-regularized feature extractor directly.

Bilinear Representations. Bilinear Pooling [23] and its variants [5, 9, 15, 16, 47] play an essential role in fine-grained classification. Recent works [50, 45, 51] also utilize such co-occurrence statistics to address few-shot learning and demonstrate positive improvements. However, standard BP [23] model suffers from high memory consumption. In-

stead, we leverage a low-rank variant [16] that shows not only low memory consumption but also better performance compared to the standard one.

3. Proposed Approach

3.1. Preliminaries

Problem Definition. There are usually two separated learning stages in the few-shot learning for knowledge transfer, which correspond to three disjoint datasets: a base set $D_b = \{(x, y)\} \subset X \times Y_b$ for training, a validation set $D_v = \{(x, y)\} \subset X \times Y_v$ for model selection, and a novel set $D_n = \{(x, y)\} \subset X \times Y_n$ for testing, where (x, y) is a pair of a sample and the corresponding label. Label spaces in the three datasets are pairwise disjoint. In the meta-training stage, a few-shot learning algorithm aims to extract general knowledge from D_b that could be reused for recognizing novel classes. In the meta-testing stage, a novel N-way K-shot classification task, a small support set and a query set pair $\{S, Q\}$, is randomly sampled from the novel set D_n. Support set S contains N different classes with K samples each: $S = \{S_i\}_{i=1}^{N}, |S_i| = K$. The objective is to classify each query example $x_q \in Q$ into the correct support class with the model adaptation to support set S.

Nearest-Centroid Classifier (NCC). NCC follows the nearest neighbor rule. In a novel few-shot task, a query sample x_q is compared with each class centroid c_j by a particular distance metric. The conditional probability of the query sample x_q belonging to the class k is:

$$p_\theta(y = k|x_q) = \frac{\exp\left(-d\left(f_\theta(x_q), c_k\right)\right)}{\sum_j \exp\left(-d\left(f_\theta(x_q), c_j\right)\right)}, \qquad (1)$$

where $d(.,.)$ denotes a distance function, $f_\theta()$ is the feature extractor function usually regard as a crucial component for few-shot classification. In order to get the feature extractor by training with the base set D_b, there are two alternative pre-training manners. One [39, 41] utilizes episodic training mechanism with the spirit of meta-learning. The other [11, 32, 12] directly proceeds to a regular supervised training routine with a learnable classification layer[1]. To generate exacter class centroids of novel categories, [11, 12] further train a weight generator with an additional episodic training stage. In contrast, a more general way to build the nearest centroid classifier is calculating centroids by the mean vector of the embedded support features belonging to the same class:

$$c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i = k) \in S_k} f_\theta\left(x_i\right) \qquad (2)$$

Further, a query sample x_q is assigned to the class of nearest centroid according to Eq. 1.

[1]Typically, those methods use the negative cosine similarity as the distance metric in the target classier, and we also regard the classifier weights as class centroids.

3.2. Improved Feature Representations with LBP

Given an input image x, we suppose the feature map is the output of the feature extractor without global pooling as $f_\theta(x) \in \mathbb{R}^{h \times w \times c}$, where c is channels. So the feature map can be decomposed following the spatial dimension: $f_\theta(x) = [f_\theta(x)_1, \cdots, f_\theta(x)_{hw}]$. Original low-rank bilinear pooling(LBP) [15] utilizes Hadamard product for an efficient attention mechanism of multi-modal learning. We apply the essential LBP representation for facilitating the few-shot classification. According to HBP [47], the low-rank bilinear pooling can be written as:

$$f_B(x) = \sum_{l=1}^{hw} U^T f_\theta(x)_l \circ V^T f_\theta(x)_l \qquad (3)$$

where $U \in \mathbb{R}^{c \times d}$ and $V \in \mathbb{R}^{c \times d}$ are projection matrices for dimension reduction, so the feature dimension can be manually decreased to d. As shown in Fig. 2(b), the architecture of LBP in our application consists of two 1×1 convolutions followed by a Hadamard product and a global average pooling. Following BCNNs [23], we also add a l_2 normalization at the end.

3.3. Class-wise Metric Scaling(CMS)

Class-wise Metric Scaling is the key novelty of our approach for improving feature representations and task adaptation in the meta-training and meta-testing stage, respectively. In the next two subsections, we elaborate on this mechanism for two stages in more detail.

3.3.1 CMS for Embedding Learning

In the meta-training stage, we adopt the conventional supervised training procedure following [11, 32], which pretrains the feature extractor combined with a distance metric based classification layer that we regard as a set of learnable class centroids. Formally, to perform CMS, we define a learnable metric scaler vector $\mathbf{s}^b = \{s_i\}_{i=1}^{N_b} \in R_+^{N_b}$ that each component corresponds to a base class centroid in $\mathbf{c}^b = \{c_i^b\}_{i=1}^{N_b} \in R^{N_b \times d}$, where N_b denotes the number of base classes and d denotes the feature dimension. $f_B()$ represents feature extractor. The training procedure is to minimize the following classification loss over parameters $\{B, \mathbf{c}^b, \mathbf{s}^b\}$ with the whole base set D_b.

$$\min_{B, \mathbf{c}^b, \mathbf{s}^b} \mathbb{E}_{(x,y) \in D_b} - \log\left(\frac{\exp\left(-s_y^b d\left(f_B(x), c_y^b\right)\right)}{\sum_j \exp\left(-s_j^b d\left(f_B(x), c_j^b\right)\right)}\right), \qquad (4)$$

where $d(.,.)$ is the distance function, *i.e.*, Euclidian distance or negative cosine similarity. We target obtaining a more discriminative and transferable feature representation.

Discussion. Scaling the cosine similarity in the softmax loss is a practical skill for improving the embedding learning

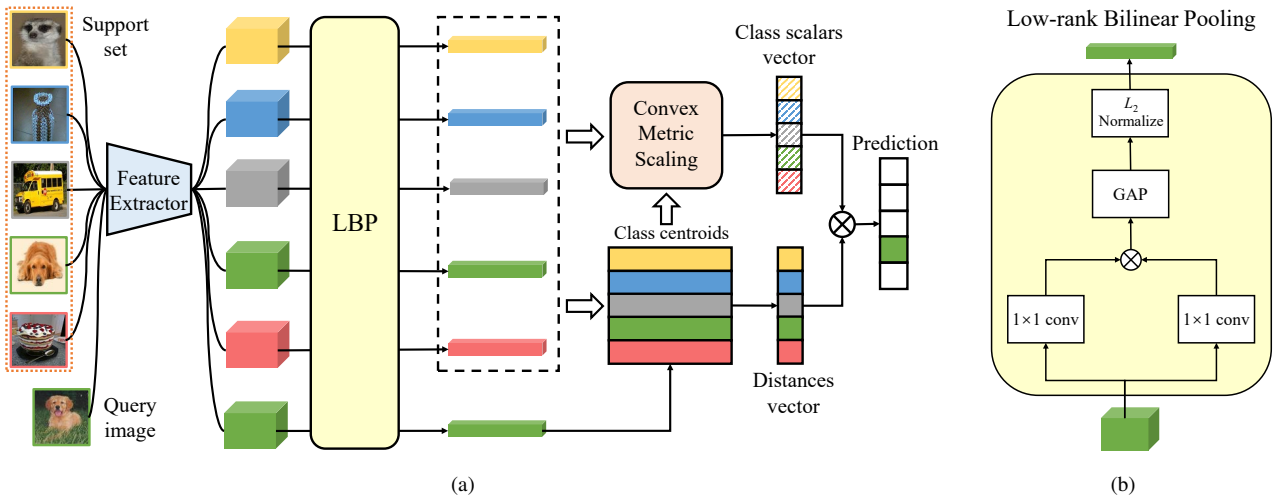


Figure 2. **(a)** Overview of our CMS model. **(b)** Architecture of LBP that we adapted. "1×1 **conv**": Convolutional layer with 1×1 kernel. "**GAP**": Global average pooling. "L_2 **Normalize**": L_2 normalization layer. (Best viewed in color)

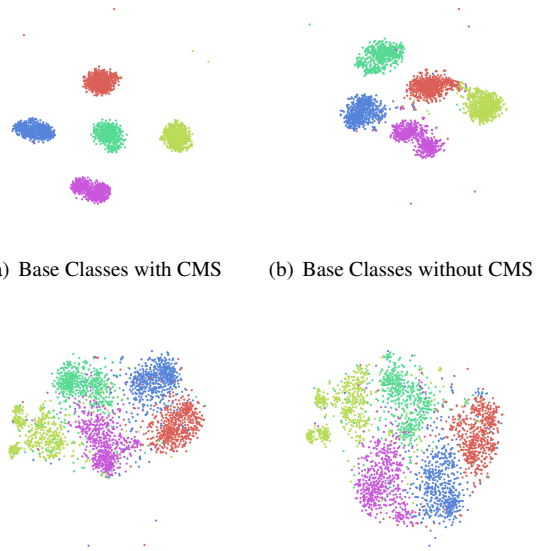
in common classification tasks [34, 43, 44, 52] and recently also applied in the few-shot learning [32, 31, 11, 4], where all the similarity scores share an equal scale factor s. In contrast, we develop a class-wise scaling manner with the motivation of Fig.1, and the metric is not limited to the cosine similarity. Besides, we have proved that leveraging CMS is equivalent to introducing a more flexible class representation by considering intra-class variances in Section 3.4.

Qualitative visualization with CMS. Overall, the positive effects of those scaling skills for embedding learning can be attributed to better feature discriminability learned from base classes and transferability to novel classes. To empirically figure out the superiority of CMS compared to the *equal metric scaling* for the embedding learning, we visualize learned feature distributions by t-SNE [28]. It can be explicitly observed from Fig.3 that the learned feature spaces **with** the CMS mechanism are intrinsically better than the ones learned **without** CMS (i.e., an equal scaling case). Concretely, CMS encourages better intra-class compactness and inter-class separability for both base and novel classes.

3.3.2 Optimal CMS for Novel Few-shot Tasks

Unlike prior metric scaling methods that set an equal scalar to all categories only benefit embedding learning, our CMS can utilize task-specific knowledge to improve generalization during meta-testing further. Instead of involving a parametric adaptation module with an additional episodic training stage, our CMS is performed directly on given novel few-shot tasks with a convex optimization problem.

Given a novel few-shot task, CMS aims at optimizing the posterior probability on the support set S, leading to optimal scalar parameters $\mathbf{s}^n = \{s_i^n\}_{i=1}^N$. With the frozen feature extractor $f_B()$, we denote samples in support set already passed



(a) Base Classes with CMS (b) Base Classes without CMS

(c) Novel Classes with CMS (d) Novel Classes without CMS

Figure 3. **t-SNE Feature Visualization Learned with/without CMS**. The feature representations are learned on *mini*-ImageNet. Five base classes in (a) and (b) are *house finch*, *robin triceratops*, *green mamba*, *harvestman*. Five novel classes in (c) and (d) are *nematode*, *king crab*, *golden retriever*, *malamute*, *dalmatian*.

from f_B as a feature point set $S' = \{(z_i, y_i)\}_{i=1}^{K \times N}$. Following the NCC classifier, the class centroids $\mathbf{c}^n = \{c_i^n\}_{i=1}^N$ are calculated by the average of corresponding support features with Eq. 2. The objective is to find optimal parameters $\mathbf{s}^n$ that maximize the likelihood on the support point set S', which is equivalent to minimizing the negative log-likelihood as follow:

$$J(\mathbf{s}^n) = \frac{1}{NK} \sum_{(z_i,y_i)\in S'} -\log\left(\frac{\exp\left(-s^n_{y_i} d\left(z_i, c^n_{y_i}\right)\right)}{\sum_j \exp\left(-s^n_j d\left(z_i, c^n_j\right)\right)}\right), \tag{5}$$

where distance $d\left(z_i, c_j\right)$ is a constant for the frozen feature space and we re-denote it as $d_{i,j}$. By expanding the log-likelihood function and adding a ℓ_2 regularization term, the total objective function is to minimize the following optimization problem among the variable $\mathbf{s}^n$:

$$\begin{aligned} &\min_{\mathbf{s}^n} \sum_{(z_i,y_i)\in S'} \left[d_{i,y_i} * s^n_{y_i} + \log \sum_j e^{-d_{i,j} * s^n_j} \right] + \beta \left\|\mathbf{s}\right\|_2, \\ &s.t. \quad s^n_i > 0 \quad i \in \{1, \ldots, N\} \end{aligned} \tag{6}$$

where β is a non-negative weighting coefficient. The optimization problem has two key characteristics as follows.

Convexity. Besides the regularization term, the objective optimization function contains two parts, the first part $d_{i,y_i} * s^n_{y_i}$ is a linear combination of the scaler vector $\mathbf{s}^n$, and the second part is a log-sum-exp function with another linear combination of scaler vector. The log-sum-exp function is convex and nondecreasing in its every argument. According to the vector composition rule [2], Eq.6 is convex.

Low-dimensionality. It is worth noting that only N (usually is 5) variables in the optimization problem. Benefiting from the low-dimensionality, we apply a second-order optimization algorithm BFGS as the solver. The average value of metric scalars $\mathbf{s}^b$ of base classes is used to initialize $\mathbf{s}^n$ for fast convergency.

The overview of the few-shot testing procedure has been illustrated in figure 2(a). Finally, for a given query sample x_q, the metric-scaled prediction with optimal scalars is:

$$\hat{y}_q = \arg \max_{j\in\{1,\ldots,N\}} -s^n_j d\left(f_B(x_q), c^n_j\right) \tag{7}$$

3.4. Interpretation as Gaussian Mixture Model

For a given N-way classification task, we can assume the feature point set follows a Gaussian Mixture distribution: $f(x) = \sum_{i=1}^{N} \alpha_i \phi(x; \mu_i, \Sigma_i)$, where each class distribution is regarded as an individual Gaussian function ϕ, and α_i is the prior probability belong to the class i with $\sum_i \alpha_i = 1$. For simplicity, we consider spherical Gaussian distributions with a particular uniform variance $\Sigma_i = \sigma^2 I \in R^{d\times d}$, and the densities of the form as:

$$\phi(x; \mu, \sigma) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp(-\frac{\parallel x - \mu \parallel^2}{2\sigma^2}) \tag{8}$$

Conditional probability of a feature $x \in R^d$ belong to the ground truth class k can be derived as follows:

$$\begin{aligned} p(y=k|x) &= \frac{\alpha_k \phi(x; \mu_k, \sigma_k)}{\sum_i \alpha_i \phi(x; \mu_i, \sigma_i)} = \frac{\frac{\alpha_k}{\sigma^d_k} \exp(-\frac{\|x-\mu_k\|^2}{2\sigma^2_k})}{\sum_i \frac{\alpha_i}{\sigma^d_i} \exp(-\frac{\|x-\mu_i\|^2}{2\sigma^2_i})} \\ &\overset{\substack{\alpha_i=\frac{\sigma^d_i}{\sum_{i'}\sigma^d_{i'}} \\ s_i=\frac{1}{2\sigma^2_i}}}{=\!=\!=\!=\!=\!=} \frac{\exp(-s_k \parallel x - \mu_k \parallel^2)}{\sum_i \exp(-s_i \parallel x - \mu_i \parallel^2)} \end{aligned} \tag{9}$$

Thus, the posterior probabilities of CMS and GMM are equivalent with a condition $\alpha_i = \frac{\sigma^d_i}{\sum \sigma^d_{i'}}$, in which case the metric scalar s_i is equal to $\frac{1}{2\sigma^2_i}$ with the same class index i. Now, our motivation in Fig. 1 that metric scalars correspond to intra-class variances has been proved, i.e., **the class with a larger variance σ_i shall be assigned with a smaller scaler value s_i** after optimization. As a result, involving the CMS mechanism is equivalent to considering class variances. In addition, extending the vanilla NCC into a CMS-based one essentially allows for more flexible class representations, which implicitly captures the intra-class variances. With classes sharing the same scalar/variance(s/σ), the CMS classifier can be further simplified to the NCC.

3.5. Extension to Semi-supervised Modes

Semi-supervised learning improves the model generalization by an unlabeled auxiliary set. We consider two kinds of semi-supervised learning modes for facilitating few-shot classification, i.e., semi-supervised transductive (**SS-T**)[2] [26, 33] and semi-supervised inductive (**SS-I**) [36, 21] modes. SS-T mode directly utilizes the samples from the query set Q as the auxiliary dataset, yet the auxiliary set apart from support and query sets is sampled in SS-I mode.

We simply leverage pseudo-labeling [17] and cherry-picking [21] mechanisms to uniformly address both semi-supervised few-shot learning tasks. Our goal is to pick the samples with high confidence for augmenting the support set S. So meta-testing optimization procedures remain the same as in the standard few-shot learning setting.

Concretely, each unlabeled sample in the auxiliary set is pseudo-labeled with the non-scaled distance metric first, and the probability score is also regarded as the confidence coefficient of the pseudo-labeling. Then the part of samples with higher confidence is added to the support set, in which case we set a hyper-parameter τ as the proportion of cherry-picking.

4. Experiments

We evaluate CMS on standard, cross-domain, and semi-supervised few-shot classification benchmarks. Also, exten-

[2]With the perspective of the unlabeled dataset, transductive inference can be regarded as a special case of semi-supervised learning.

sive ablation studies and analyses are presented to validate the effectiveness of each component of our approach.

4.1. Experimental setups

Datasets. For standard few-shot classification, we evaluate on two canonical datasets: *mini*-ImageNet [41] and tieredImageNet [37]. For the cross-domain scenario, we evaluate our approach following a recently proposed benchmark [3], where *mini*-ImageNet dataset as the seen domain for training, and CUB [42] dataset as the target domain for testing.

Backbones. We use **Conv4** [41] and **ResNet12** [31] as the feature extractors in our experiments. Detailed model descriptions are provided in Appendix A. Feature dimension d of the LBP is set as 1024 and 8192 for Conv4 and ResNet12, respectively.

Implementation details. Conventional training and testing setups are given in Appendix B. Particularly, in the training stage, we adopt conventional data augmentation protocols, including random crop, horizontal flip, and color jitter. In the testing stage, ten crops and image flip are used for both support set and query set. For support samples, the augmented samples are equally used for calculating centroids and solving the convex problem. For a query sample with its augmented variants, the probability scores are averaged for the final prediction. In all 1-shot and 5-shot experiments, we randomly sample 5000 few-shot tasks with 15 queries per class and report the average accuracy with a corresponding 95% confidence interval. In the SS-I inference mode, an auxiliary set containing 50 images per class is sampled. Hyper-parameter and model selection are conducted according to the mean accuracy on the validation set.

4.2. Experimental results

4.2.1 Performance on Standard Few-shot Benchmarks

Table 1 summarizes the results of the 5-way 1-shot/5-shot classification on *mini*-ImageNet and tieredImageNet datasets. Overall, our CMS achieves state-of-the-art results with relatively low variance intervals.

Distance Metric Comparison. We first perform CMS based on two regular metric distances, *i.e*., Euclidean distance and cosine similarity. As can be observed, cosine similarity based CMS always outperforms Euclidean distance-based one, which is consistent with the choice in other NCC based approaches [4, 6, 7, 24]. So in the subsequent experiments and discussions, we focus on cosine similarity based CMS.

Backbone comparison. It is evident that a deeper feature extractor backbone significantly boosts the few-shot performance. Moreover, LBP improves the performance significantly, especially for the Conv4 based models, which also reveals that the Conv4 model capacity is insufficient to distill knowledge from both datasets. However, We must emphasize that performance gains by LBP are not only due to the increase of model parameters, as the parameter efficacy of LBP is analyzed in Section 4.4.

Results on *mini*-ImageNet. Our CMS achieves the best 1-shot and 5-shot classification results on both shallow Conv4 and deep ResNet12 backbones, and even outperforms methods with deeper backbones (such as ResNet18 and WRN28). To be specific, in 1-shot setting, our CMS achieves the accuracy of 66.64% which outperforms the prior leading method *centroid alignment* [1] by 0.72%. In 5-shot setting, CMS obtains the accuracy of 83.63% that surpasses the method *S2M2*$_R$ [29]. Moreover, CMS also shows its superiorities over other distance metric based methods [10, 31, 4, 19, 39, 40, 41, 49].

Results on tieredImageNet. CMS also achieves state-of-the-art results on tieredImageNet. To be specific, in 5-shot setting, our CMS achieves the accuracy of 87.66% outperforms the prior leading method *centroid alignment* [1] by 1.1%. Although *centroid alignment* [1] performs better than ours on the 1-shot setting, it uses a much deeper backbone WRN28. In addition, CMS outperforms *centroid alignment* with a ResNet18 backbone by 4.2% when compared under models with similar capacity.

4.2.2 Performance on Cross-domain Benchmark

Cross-domain is a more challenging scenario for few-shot learning. We conduct experiments to validate the effectiveness of CMS following the recent benchmark [3]. Table 2 shows our experimental results compared with prior approaches. It is obvious that the CMS outperforms all previous methods. Concretely, CMS is significantly better than prior SOTA method *centroid* [1] by 9.8% and 8.2% for 5-way 1-shot and 5-shot settings, respectively. It indicates that our approach also owns the high capacity to mitigate the domain shift.

4.2.3 Semi-supervised Few-shot Learning

We evaluate the performance of CMS for SS-T and SS-I modes (introduced in Section 3.5) on *mini*-ImageNet dataset, and results are shown in Table 3. For two types of semi-supervised learning, we can observe further large improvement margins compared with the standard mode of Table 1, which indicates auxiliary data immensely enrich the class distribution information, and make it possible for CMS to improve model generalization. For example, CMS obtains the accuracy of 78.34%/87.53% in 1-shot/5-shot SS-I mode, which improves 11.7%/3.9% compared to the standard few-shot mode.

For the SS-T mode, CMS surpasses the SOTA method DPGN [46], especially in the 1-shot setting. In the 5-shot SS-I mode, CMS surpasses *TransMatch* [48] by 6.3%. In

Table 1. **Comparison to prior methods on ImageNet derivatives**. Best results are highlighted. *Results from [18].

		***mini*-ImageNet 5-way**		**tieredImageNet 5-way**	
Method	**Backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
ProtoNet* [39]	Conv4	49.42±0.78	68.20±0.66	53.31±0.89	72.69±0.74
RelationNet* [40]	Conv4	50.44±0.82	65.32±0.70	54.48±0.93	71.32±0.78
MatchingNet* [41]	Conv4	43.56±0.84	55.31±0.73	-	-
CMS(Euclidean)	Conv4	56.31±0.28	76.60±0.22	60.70±0.31	79.56±0.23
CMS(Cosine, w/o LBP)	Conv4	54.41±0.28	71.82±0.22	57.35±0.31	74.18±0.25
CMS(Cosine)	Conv4	58.99±0.28	77.10±0.22	62.80±0.31	79.72±0.24
TADAM [31]	ResNet12	58.50±0.30	76.70±0.30	-	-
MetaOptNet [18]	ResNet12	62.64±0.62	78.63±0.46	65.99±0.72	81.56±0.53
Meta-Baseline [4]	ResNet12	63.17±0.23	79.26±0.17	68.62±0.27	83.29±0.18
DeepEMD [49]	ResNet12	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58
CTM [19]	ResNet18	64.12±0.82	80.51±0.13	68.41±0.39	84.28±1.73
centroid alignment [1]	ResNet18	59.88±0.67	80.35±0.73	69.29±0.56	85.97±0.49
LEO[38]	WRN28	61.76±0.08	77.59±0.12	66.33±0.05	81.44±0.09
CC+rot[10]	WRN28	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36
S2M2$_R$ [29]	WRN28	64.99±0.18	83.07±0.13	-	-
Neg-Cosine [24]	WRN28	61.72±0.81	81.79±0.55	-	-
centroid alignment [1]	WRN28	65.92±0.60	82.85±0.13	**74.40±0.68**	86.61±0.59
CMS(Euclidean)	ResNet12	63.22±0.28	82.14±0.20	70.94±0.31	86.87±0.20
CMS(Cosine, w/o LBP)	ResNet12	64.78±0.28	82.38±0.18	71.13±0.31	86.05±0.22
CMS(Cosine)	ResNet12	**66.64±0.28**	**83.63±0.18**	73.48±0.31	**87.66±0.20**

Table 2. **Comparison to prior methods on Cross-domain Benchmark**. Models is trained with *mini*-ImageNet dataset and evaluated on CUB dataset. Our CMS uses ResNet12 backbone.

Method	**5-way 1-Shot**	**5-way 5-Shot**
ProtoNet [39]	-	62.02±0.70
MAML [8]	-	51.34±0.72
RelationNet [40]	-	57.71±0.73
Baseline [3]	-	65.57±0.70
Diverse-20 [6]	-	66.17±0.55
Neg-Softmax [24]	-	69.30±0.73
centroid [1]	47.25±0.76	72.37±0.89
CMS(w/o LBP)	54.46±0.29	75.49±0.23
CMS	**57.02±0.29**	**80.56±0.21**

Table 3. **Results of semi-supervised modes on *mini*-ImageNet dataset**. Two evaluation setups: Semi-supervised transductive (**SS-T**) and Semi-supervised inductive (**SS-I**) modes.

Mode	**Method**	**5-way 1-shot**	**5-way 5-shot**
SS-T	TEAM [33]	60.07±-	75.90±-
	TPN [26]	59.46±-	75.65±-
	BD-CSPN [25]	70.31±0.93	81.89±0.60
	DPGN [46]	67.77±0.32	84.60±0.43
	CMS(Our)	**74.54±0.32**	**86.64±0.17**
SS-I	Soft k-Means [36]	51.52±0.36	70.25±0.31
	LST [21]	70.1±1.9	78.7±0.8
	TransMatch [48]	63.02±1.07	81.19±0.59
	CMS(Our)	**78.34±0.31**	**87.53±0.16**

the 1-shot SS-I mode, CMS improves accuracy by 8.2% compared to SOTA method LST [21]. Particularly, only 50 images are sampled as the auxiliary set in our experiments while it is 100 in LST.

4.3. Full ablation study

We conduct detailed ablation studies to explore the contributions of three components: LBP(low-ranking bilinear pooling), CMS-train (CMS for embedding learning), and CMS-test (CMS for task-adaptation). The results are shown in Table 4 based on three experimental scenarios: *mini*-ImageNet with Conv4, *mini*-ImageNet with ResNet12, and cross-domain with ResNet12.

Baseline. Our baseline model, the cosine similarity based NCC classifier, already achieves competitive performance compared with most previous methods, which can be attributed to data augmentation protocols in both training and testing stages, e.g., mean accuracy on 5-shot *mini*-ImageNet setting are 80.78% and 69.64% with ResNet12 and Conv4 backbones, respectively. The same observations also hold for the cross-domain scenario.

Effect of LBP. LBP provides significant improvements over the baseline model. For example, performance gains

	CMS		*mini*(Conv4)		*mini*(ResNet12)		*cross*(ResNet12)	
LBP	**train**	**test**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
			51.55	69.64	62.38	80.78	53.75	74.07
✓			56.68	75.12	64.55	81.83	56.37	78.03
✓	✓		59.00	76.55	66.57	83.05	56.93	79.52
✓	✓	✓	58.99	77.10	66.64	83.63	57.02	80.56

Table 4. **Ablation studies on three few-shot classification scenarios**. The baseline model (first row) is the cosine-similarity based NCC. *mini* denotes the *mini*-ImageNet dataset, and cross denotes the cross-domain benchmark.

2.2%/1.1% on 1-shot/5-shot *mini*-ImageNet with ResNet12. For the Conv4 backbone, the performance gains are more prominent. The improvements are up to 5.1% and 5.5% for 1-shot and 5-shot settings, respectively.

Effect of CMS for embedding learning. We further study the effects of CMS in the training stage, which introduces the learnable class metric scalars and provides significant improvements compared to an equal-scaled NCC, e.g., accuracy gains 2.0%/1.2% for ResNet12 on the 1-shot/5-shot *mini*-ImageNet setting.

Effect of CMS for task-adaptation. In the testing stage, CMS involves the optimization problem for producing optimal metric scalars. For 5-shot setting, performance improvements are 0.6%, 0.6%, 1.0% on three scenarios, respectively. No obvious improvements are observed for 1-shot learning since only one sample per class cannot effectively represent the intra-class variances.

Overall, taking the 5-shot setting as an example, we observe a relative accuracy gain with 7.5%, 2.9%, and 6.5% for three scenarios, respectively.

4.4. Further Analysis and Discussion

Accuracy w.r.t. the number of unlabeled samples. Varying the number of unlabeled samples, we report mean accuracies of both types (SS-I and SS-T) of semi-supervised modes in Table 5. As the number of unlabeled samples increases from 0 to 70, the mean accuracy of both modes continuously improves.

Table 5. **Mean accuracies (%) of two semi-supervised modes by varying the number of unlabeled samples**.

Number	0	15	30	50	70
SS-I 1-shot	66.64	76.46	77.56	78.34	78.77
SS-I 5-shot	83.63	86.50	87.29	87.53	87.83
SS-T 1-shot	66.64	74.54	76.64	77.70	78.10
SS-T 5-shot	83.63	86.64	87.27	87.68	87.74

Parameter efficiency of LBP. One may raise a concern that LBP boosts few-shot results due to increasing the model parameters, especially on the shallow Conv4 backbone. To address the concern, we further construct a backbone by extending Conv4 with two additional convolutional layers, naming it as Conv6, which is slightly larger than *Conv4+LBP*. The NCC performance based on those backbones is compared in Table 6. The *Conv4+LBP* significantly surpasses the *Conv6* with fewer parameters, which indicates LBP is very parameter-efficient in extracting discriminative features for addressing few-shot classification.

Table 6. **NCC performance with different backbones**. Models are trained on the *mini*-ImageNet dataset.

Backbone	**Size** (MB)	**1-Shot**	**5-Shot**
Conv4	0.43	51.55±0.28	69.64±0.24
Conv6	0.96	55.67±0.28	73.85±0.23
Conv4+LBP	0.93	56.68±0.28	75.12±0.22

Whether the CMS parameters relate to per-class variances? To quantitatively verify the relationship of intra-class variances and learned metric scalars in Section 3.4, we use the *Pearson correlation coefficient* for measuring the correlation. At the end of embedding learning, we directly calculate class variances with $\sigma_k^2 = \sum_i \parallel x_i - c_k \parallel^2$, where c_k denote a class centroid and x_i is the sample belong to the class. For 64 base classes in the *mini*-ImageNet, the Pearson correlation coefficient of variance vector $\mathbf{V}_\sigma = \{\sigma_i\}_{i=1}^{64}$ and the learned scalar vector $\mathbf{s}^b = \{s_i\}_{i=1}^{64}$ is -0.87, which indicates a high-negative correlation and justifies the demonstration in Section 3.4, *i.e.*, a class with larger intra-variance corresponds to a smaller metric scalar.

5. Conclusion

In this paper, we propose a Class-wise Metric Scaling(CMS) mechanism for further improving the few-shot classification performance based on the NCC classifier. In particular, metric scalars are set as learnable parameters in the training stage, which results in more discriminative and transferable feature representations. As for testing, based on the maximum likelihood technique, a convex optimization program has been introduced to generate optimal metric scalars for few-shot tasks. Moreover, CMS can be interpreted as a special case of the Gaussian mixture model mathematically. Besides, we also incorporate a low-ranking bilinear pooling layer for improving representation capacity. Extensive experiments among multiple backbones and datasets demonstrate consistent improvements compared to prior state-of-the-art methods.

References

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Associative alignment for few-shot image classification. In *European Conference on Computer Vision (ECCV)*, 2020.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[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 *International Conference on Learning Representations*, 2019.

[4] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning, 2020.

[5] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2930, 2017.

[6] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Diversity with cooperation: Ensemble methods for few-shot classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[7] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting relevant features from a multi-domain representation for few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2020.

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

[9] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Perez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Ross Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

[15] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations (ICLR)*, 2017.

[16] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

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

[19] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Gao Yang, and Jiebo Luo. Distribution consistency based covariance metric networks for few-shot learning. In *AAAI*, 2019.

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

[22] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[24] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2020.

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

[26] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations*, 2019.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

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

[30] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999, 2018.

[31] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems 31*, pages 721–731. 2018.

[32] Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[34] Rajeev Ranjan, Carlos Domingo Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *CoRR*, abs/1703.09507, 2017.

[35] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2016.

[36] 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 *International Conference on Learning Representations ICLR*, 2018.

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

[38] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019.

[39] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087. 2017.

[40] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[41] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[43] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

[44] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] Davis Wertheimer and Bharath Hariharan. Few-shot learning with localization in realistic settings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu. Dpgn: Distribution propagation graph network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13390–13399, 2020.

[47] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[48] Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12856–12864, 2020.

[49] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12203–12213, 2020.

[50] Hongguang Zhang and Piotr Koniusz. Power normalizing second-order similarity network for few-shot learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1185–1193, 2019.

[51] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[52] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.