ReMarNet: Conjoint Relation and Margin Learning for Small-Sample Image Classification

Xiaoxu Li, Liyun Yu, Xiaochen Yang, Zhanyu Ma, *Senior Member, IEEE,* Jing-Hao Xue, *Member, IEEE,* Jie Cao, Jun Guo

***Abstract*—Despite achieving state-of-the-art performance, deep learning methods generally require a large amount of labeled data during training and may suffer from overfitting when the sample size is small. To ensure good generalizability of deep networks under small sample sizes, learning discriminative features is crucial. To this end, several loss functions have been proposed to encourage large intra-class compactness and inter-class separability. In this paper, we propose to enhance the discriminative power of features from a new perspective by introducing a novel neural network termed Relation-and-Margin learning Network (ReMarNet). Our method assembles two networks of different backbones so as to learn the features that can perform excellently in both of the aforementioned two classification mechanisms. Specifically, a relation network is used to learn the features that can support classification based on the similarity between a sample and a class prototype; at the meantime, a fully connected network with the cross entropy loss is used for classification via the decision boundary. Experiments on four image datasets demonstrate that our approach is effective in learning discriminative features from a small set of labeled samples and achieves competitive performance against state-of-the-art methods. Code is available at https://github.com/liyunyu08/ReMarNet.**

***Index Terms*—Small-sample learning, Deep neural network, Relation learning, Discriminative feature learning.**

I. Introduction

Deep learning has achieved state-of-the-art results in various visual tasks, including image and video classification [1], [2], [3], [4], [5], object recognition [6], [7], and semantic segmentation [8]. However, its superior performance heavily relies on a large number of labeled training samples, which are difficult to acquire in many cases, thus severely limiting its application in real life. In addition, when the size of training set is small, the deep model will inevitably suffer from overfitting as the network architecture goes deeper. Hence, how to avoid overfitting and obtain a model with good generalizability under the condition of small sample sizes is a great challenge.

Many methods have been proposed to reduce overfitting in the case of small sample sizes, which can be mainly divided into data enhancement [9], domain adaptation [10], [11], regularization [12], [13], network ensemble [14], and feature extraction [15], [16]. Recently, in the field of feature extraction, there has been a growing number of research on learning discriminative features as a way of preventing overfitting in neural networks. The fundamental pipeline is to optimize a loss function toward better intra-class compactness and inter-class separability. However, most existing methods make assumptions about the type of metric or data distribution beforehand, and these assumptions limit the adaptability of these methods to different tasks.

In this paper, we propose a new method for learning the discriminative features and performing classification. Our motivation derives from two aspects. Firstly, as illustrated in Figure 1, intuitively, if the features can support both the classification paradigm based on learning decision boundary between different classes and the classification paradigm through comparing the similarity to class prototypes, the discriminability of features will be enhanced. Secondly, inspired by the recognition mechanism of human beings, these two classification paradigms are often considered jointly to identify the category of an unseen object; that is, the prediction will be made by considering the outcomes of two paradigms jointly. Building on these two aspects, we propose a *Relation-and-Margin learning neural Network* (ReMarNet) for small-sample image classification, which could perform feature learning and classification from two perspectives jointly.

To implement our proposal, the ReMarNet consists of a feature embedding module and a classification module. The classification module comprises two branches. One branch is constructed from the relation network [17], which reduces the distance between each sample and its corresponding class prototype and thereby improves the intra-class compactness. The other branch is a two-layer fully connected network with the cross-entropy loss, which guarantees the prediction accuracy. The network is trained in an end-to-end fashion so as to learn the discriminative features that can conjointly learn the satisfactory separation margin and prototype similarity for better small-sample classification. The final prediction is produced by assembling the outputs of two network branches for better generalization.

To investigate the effectiveness of the proposed method, we

This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303300 and under Subject II No. 2019YFF0303302, in part by the National Natural Science Foundation of China (NSFC) under Grant 61906080, 61763028, 61773071, 61922015, and U19B2036, in part by the National Science and Technology Major Program of the Ministry of Science and Technology under Grant 2018ZX03001031, in part by the Beijing Academy of Artificial Intelligence (BAAI) under Grant BAAI2020ZJ0204, in part by the Beijing Nova Programme Interdisciplinary Cooperation Project under Grant Z191100001119140, in part by the Key Program of Beijing Municipal Natural Science Foundation under Grant L172030, in part by the Hong-liu Outstanding Youth Talents Foundation of Lanzhou University of Technology.

Corresponding author: Zhanyu Ma (email: mazhanyu@bupt.edu.cn)

X. Li, L. Yu and J. Cao are with the School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China.

X. Li, Z. Ma and J. Guo are with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China.

X. Yang and J.-H. Xue are with the Department of Statistical Science, University College London, London, WC1E 6BT, U.K.

X. Li, L. Yu and X. Yang contribute equally.

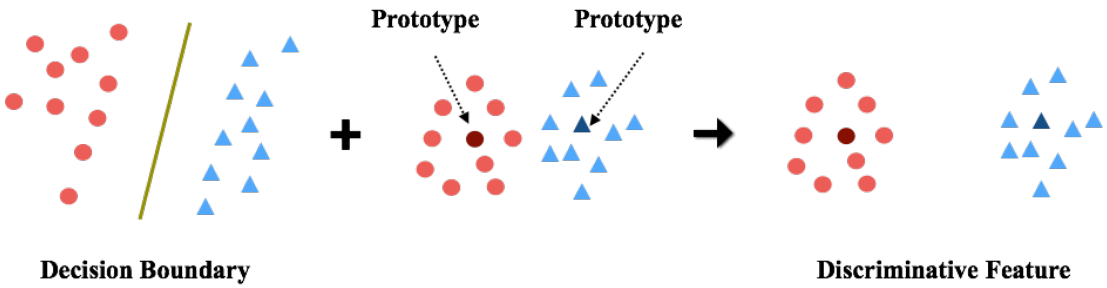


Fig. 1. The motivation of the proposed Relation-Margin learning neural Network (ReMarNet). Taking an example of binary classification: the round and triangle points represent two classes, respectively. In each class, the sample that an arrow points to denotes the prototype of the class. A green line denotes the decision boundary of two classes. The discriminability of features will be enhanced if they could excel in both classification paradigms, i.e. the paradigm based on learning decision boundary and the paradigm through comparing the similarity to class prototypes.

conduct experiments on four real image datasets for small-sample classification. Results suggest that assembling two network structures is superior to using a single-branch network and it achieves the state-of-the-art performance compared with existing loss-based methods and two ensemble methods. Our contributions are twofold:

- To the best of our knowledge, we propose the first network of integrating two kinds of classification mechanisms, i.e. the classification mechanism based on prototype similarity and the classification mechanism based on decision boundary, termed Relation-and-Margin learning neural network (ReMarNet), for classification with a small number of training samples. It allows for classification separately or conjointly, as preferred by practitioners.
- Experimental results on four small-sample image datasets show that, compared with the latest work on learning discriminative features via loss functions, our method can obtain more discriminative features and superior performance.

II. Related Work

Small-sample learning has received considerable attention in the machine learning field. One category of small-sample learning is few-shot learning. The difference between few-shot learning and the general small-sample learning lies in the evaluation procedure. In few-shot learning [18], [19], the evaluation procedure averages out accuracy over many episodes. Each episode performs a C-class classification task, and each class includes K labeled samples; C and K are fixed constants. In the general small-sample learning, the number of classes is determined by the dataset and the number of labeled samples can be unequal. This paper focuses on the general small-sample classification of image data; for few-shot learning, we refer interested readers to [20].

The small sample size poses a challenge to deep learning methods, as they are easy to overfit when the model goes deeper. Data enhancement and domain adaptation methods are proposed to alleviate this problem through increasing the number of training samples. For example, the data-enhanced GAN model can automatically learn to augment training data [9]. The work in [21] proposes a novel way of transferring the data transformation mode of the base class to generate samples in new categories. Regularization is another widely adopted technique for mitigating overfitting of training networks under small samples. Examples include norm-based constraints [22], dropout [23], [12], early stopping [24], noise robustness [25], adversarial training [26] and multi-task learning [27]. Assembling multiple networks is known to yield more accurate and robust predictions than using a single network. To avoid high computational cost ensued from training multiple networks, Snapshot ensembling [14] and temporal ensembling [28] have been proposed, both of which combine multiple outputs obtained from a single training of the network.

Another group of methods focus on learning discriminative features. The pioneering work of [15] introduces the triplet loss to separate a positive pair (two matching samples) from a negative one (non-matching samples) by a distance margin in the Euclidean space. Compared with the Euclidean distance, large-margin loss based on the cosine similarity is more appropriate when used in conjunction with the softmax loss, which is widely used in convolutional neural networks (CNNs) and has demonstrated the capability of learning discriminative features. Building on the link between the cosine similarity and the softmax decision boundary, [29] and [30] enforce a stricter condition on the angle between the feature vector and the weight vector so as to improve the discriminative power of the softmax loss. With an L_2 normalization of the weight vector, [30] can be further regarded as imposing a margin in the angular space. Sharing a similar idea, [31] proposes the large margin cosine loss (LMCL), where the feature vector is additionally normalized and the margin constraint is placed on the cosine similarity, *i.e.*, encouraging a margin in the cosine space.

Aiming for both intra-class compactness and inter-class separation, the center loss is proposed to punish a large distance between the feature and its corresponding class center, and to jointly supervise the CNNs. It is balanced against the softmax loss via a weight parameter [32]. The idea of class centers is adopted in [33] but formulated in a different way. Instead of the Euclidean distance, the cosine similarity is calculated and normalized via the softmax function. The loss function is designed in a cross-entropy manner, thus avoiding the weight parameter. [34] assumes that all the features follow a Gaussian mixture model, and then improves the classification performance by introducing a classification margin and a

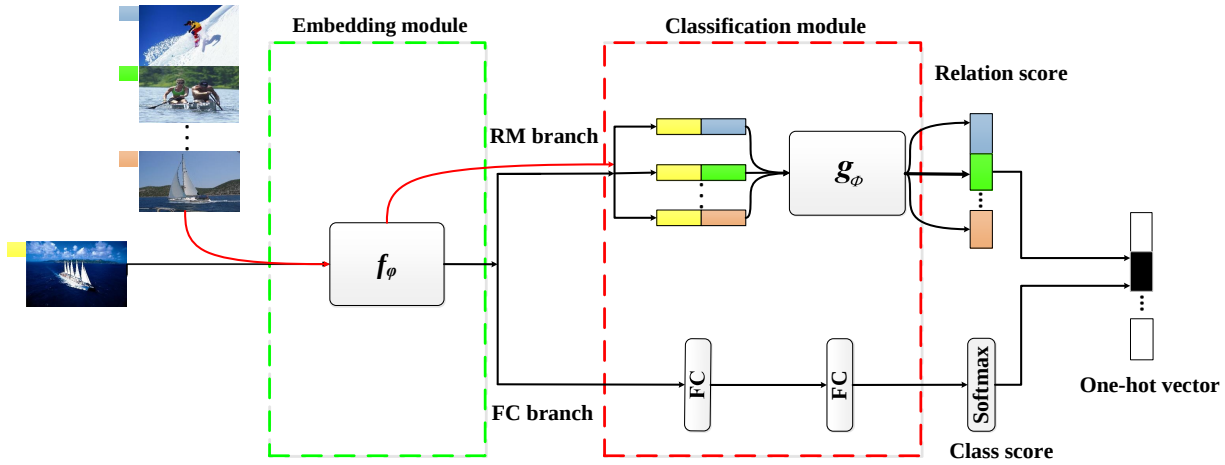


Fig. 2. Relation-and-Margin learning neural Network (ReMarNet): The network is composed of a feature embedding module f_φ and a two-branch classification module, namely the relation module (RM) branch and the fully connected network (FC) branch. Images at the top-left of the figure are the prototype samples that we select from each class; each class prototype is assigned with a different color. Given a new sample labeled in yellow, the ReMarNet predicts its class label by leveraging the scores of the two branches.

likelihood regularization term, which includes the center loss as a special case.

III. ReMarNet: Relation-and-Margin Learning Neural Network

To learn more discriminative features for small-sample classification, we conjointly enhance the intra-class compactness and enforce the inter-class separability through constructing and simultaneously learning two types of networks. Figure 2 summarizes our approach illustratively. After extracting the features via the VGG16 network [35], we shrink the distance between all the training samples and their prototype samples via the relation module (RM) to achieve the intra-class compactness, and simultaneously separate the instances from different classes via a two-layer fully connected (FC) network by the cross-entropy loss to enhance the inter-class separability. The relation score from the RM and the probability vector from the FC network will be assembled to predict the class label. Before explaining the structure of the proposed ReMarNet in detail, we first review the relation network [17], on which the RM is built.

A. Relation Network

The relation network is proposed in [17] for few-shot classification. It is constructed by two modules, namely an embedding module and a relation module.

The embedding module consists of four convolutional blocks, each of which contains a 3×3 convolution with 64 filters, a batch normalization, and a ReLU nonlinearity layer. In addition, for the first two blocks, a 2×2 max-pooling layer is placed after each block. Two outputs from the embedding module, *i.e.* two feature maps, is concatenated to construct a relation pair, which is used as the input of the subsequent relation module.

The relation module is composed of two convolutional blocks and two fully connected layers. Each convolution block consists of 3×3 convolution with 64 filters, followed by batch normalization, a ReLU activation function and a 2×2 max-pooling. The padding parameter of both blocks is set to 1. The activation functions of all the fully connected layers are ReLU except for the output layer, where a Sigmoid function is adopted in order to generate the relation scores. The input sizes of the first and second FC layers are 64 and 32, respectively, and the final output size is 1.

In summary, the input to the relation module is a concatenated feature map obtained from the embedding module, and the output is a vector of the relation score. The relation module of relation network is adopted in our proposed ReMarNet to learn the similarity between a sample and the prototype of each class.

B. Structure of ReMarNet

Consider a K-class classification task. Let $D_{train} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ denote a training dataset of N samples, where $\boldsymbol{y}_i$ is a one-hot K-dimensional vector representing the class label of $\boldsymbol{x}_i$. For later use in RM, we randomly select one sample from each class of D_{train} as prototype samples and denote them as $\{\boldsymbol{o}_j\}_{j=1}^K$.

The proposed ReMarNet comprises two parts. The first part is a feature embedding module (f_φ in Figure 2). Here we use all the convolutional blocks of the VGG16 network, which produces feature maps $f_\varphi(\boldsymbol{x}_i)$ and $f_\varphi(\boldsymbol{o}_j)$ for samples $\boldsymbol{x}_i$ and $\boldsymbol{o}_j$, respectively.

The second part is a two-branch classification module, consisting of an RM branch for optimizing the intra-class compactness and an FC branch for pushing the inter-class separability. Let $C(\cdot,\cdot)$ denote the concatenation operator of features maps. The RM branch takes a relation pair of feature maps $f_\varphi(\boldsymbol{x}_i)$ and $f_\varphi(\boldsymbol{o}_j)$, *i.e.* $C(f_\varphi(\boldsymbol{x}_i), f_\varphi(\boldsymbol{o}_j))$, as input, and learns the similarity between them through the network

g_ϕ. The output of RM is the relation score $\boldsymbol{r}_{ij}$ between $\boldsymbol{x}_i$ and $\boldsymbol{o}_j$ as

$$\boldsymbol{r}_{ij} = g_\varphi(C(f_\varphi(\boldsymbol{x}_i), f_\varphi(\boldsymbol{o}_j))) \ , \tag{1}$$

where $\boldsymbol{r}_{ij}$, ranging from zero to one, measures the similarity between the training sample $\boldsymbol{x}_i$ and the class prototype $\boldsymbol{o}_j$.

To train the RM, we compute the mean square error (MSE) between the relation score vector $\boldsymbol{r}_i$ and the ground truth label $\boldsymbol{y}_i$. The loss function for the RM branch is

$$L_{RM} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(\boldsymbol{r}_{ij} - \boldsymbol{y}_{ij})^2 \ , \tag{2}$$

where $\boldsymbol{y}_{ij}$, the jth element of $\boldsymbol{y}_i$, equals one if $\boldsymbol{x}_i$ belongs to the jth class and zero otherwise. By minimizing the RM loss, we encourage $\boldsymbol{x}_i$ to stay close to its corresponding class prototype, thereby improving the intra-class compactness.

Regarding the FC branch, we use the flattened convolutional features of training samples as input. In the last layer, the softmax activation function is used to calculate the probability $\boldsymbol{p}_i$, a K-dimensional vector where each element represents the probability that the sample $\boldsymbol{x}_i$ is assigned to each class. The FC network is trained with the following cross entropy (CE) loss:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i^T \ \log(\boldsymbol{p}_i) \ . \tag{3}$$

Minimizing the CE loss promotes learning the features that could increase the probability of assigning $\boldsymbol{x}_i$ to its ground-truth class.

Integrating the RM and FC branches, we obtain the total loss function of the proposed ReMarNet:

$$L = L_{RM} + L_{CE} \ . \tag{4}$$

In the prediction stage, we calculate the relation score between the test image and each class prototype and the probability of its belonging to each class. The test image is classified as the class with the maximum sum of the two values.

IV. Experimental Results and Analysis

Experiments in this section serve five purposes:

- To compare the proposed ReMarNet with state-of-the-art methods for the task of small-sample image classification (Sec. IV-C);
- To investigate the effect of training set size on different methods (Sec. IV-D);
- To assess the impact of different backbone networks (Sec. IV-E);
- To study the effectiveness of each branch of our network (Sec. IV-F, IV-G, IV-H);
- To evaluate the discriminative power of the learned feature embedding (Sec. IV-I).

A. Datasets

For small-sample image classification, we randomly select a subset of images from the following four datasets: LabelMe, UIUC-Sports, 15Scenes and BMW. The datasets vary in their content, number of classes and sample size.

1) LabelMe (LM) Dataset: LabelMe is a natural scene image classification dataset containing 8 classes: coast, mountain, forest, open country, street, inside city, tall buildings and highways. We randomly select 210 images from each class, of which 100 images are used to form the training set and another 100 images are used for the test set. The total number of images used in each round is 1600.

2) UIUC-Sports Dataset: UIUC-Sports contains 1578 sports scene images of 8 classes: bocce (137), polo (182), rowing (250), sailing (190), snowboarding (190), rock climbing (194), croquet (236) and badminton (200). A training set of 749 images and a test set of 749 images are randomly sampled from the entire dataset.

3) 15Scenes Dataset: 15Scenes is one of the most complete datasets for scene classification used to date in the literature, gradually built from eight classes [36] to 13 classes [37] and finally to 15 classes [38]. The total number of images is 4485 and the number per category varies between 200 and 400. The dataset is partitioned into 70% for training and 30% for testing.

4) BMW Dataset: BMW-10 is an ultra-small, fine-grained vehicle dataset comprised of 10 different types of BMW vehicles [39]. It contains a total of 512 images, and each class has around 50 images. The training and test ratio is set as 70/30.

For all datasets, we run 15 rounds of random training and test split. The mean value and standard deviation of classification accuracy are used as our evaluation criteria.

B. Methodologies and Parameter Settings

We compare the proposed ReMarNet with the baseline method, i.e., a fully connected network using the cross entropy loss (Baseline), and four state-of-the-art feature learning methods using different loss functions, namely center loss (Center) [32], L-GM loss (LGM) [34], LMCL loss (LMCL) [31], and dual loss (Dual) [40]. We also consider two ensembling networks, namely Dropout [41] and Snapshot [14].

In the proposed ReMarNet, we use the VGG16 network as our feature extractor. The number of hidden layers in the FC branch is set to 32 and details of the RM branch are explained in Sec. III-A.

All networks are trained by using the RMSprop optimizer [42] with a batch size of 32. Learning rates of 0.00001 and 0.0001 are used for the feature extraction network and the FC network, respectively. For Center, LGM and LMCL, we use the stochastic gradient descent algorithm to separately train the loss function with a learning rate of 0.01. For our method, the learning rate of the RM branch is set as 0.001. The number of epochs is 50 for all methods except for the Snapshot network where two models are trained with 50 epochs each, leading to a total of 100 epochs. LGM, LMCL and Dual involve some additional hyperparameters, which are chosen as follows: the loss weight and α in LGM are set as 0.001 and 1.5 respectively; s and m in LCML are set as the average of $\|\boldsymbol{x}\|_2$ and 0.5 respectively; the loss weight in Dual is set as 4.5.

TABLE I
COMPARISON OF THE PROPOSED REMARNET WITH STATE-OF-THE-ART METHODS. THE MEAN VALUE (MEAN) AND STANDARD DEVIATION (STD.) OF CLASSIFICATION ACCURACY ARE REPORTED WITH THE BEST RESULTS IN BOLD.

Datasets	Measure	Baseline	Center	LGM	LMCL	Dual	Dropout	Snapshot	Ours
LM	**Mean**	0.9275	0.9219	0.9136	0.9207	0.9298	0.9288	0.9271	**0.9303**
	Std.	0.0047	0.0060	0.0075	0.0155	0.0051	0.0045	0.0076	**0.0067**
UIUC	**Mean**	0.9476	0.9514	0.9492	0.9492	0.9485	0.9472	0.9437	**0.9581**
	Std.	0.0045	0.0032	0.0055	0.0052	0.0040	0.0044	0.0045	**0.0038**
15Scenes	**Mean**	0.9142	**0.9326**	0.9214	0.9243	0.9128	0.9146	0.9143	0.9310
	Std.	0.0094	**0.0037**	0.0052	0.0037	0.0052	0.0045	0.0037	0.0025
BMW	**Mean**	0.4094	0.4274	0.2329	0.4402	0.4363	0.4094	0.3936	**0.4415**
	Std.	0.0310	0.0400	0.0478	0.0354	0.0438	0.0356	0.0236	**0.0364**

TABLE II
p-VALUES OF THE WILCOXON SIGNED-RANK TEST. $*$ INDICATES THAT REMARNET IS SIGNIFICANTLY DIFFERENT FROM THE COMPARED METHOD AT THE SIGNIFICANCE LEVEL OF 5%.

Datasets	Baseline	Center	LGM	LMCL	Dual	Dropout	Snapshot
LM	0.1228	0.0006*	0.0015*	0.0153*	0.8438	0.6374	0.3107
UIUC	0.0007*	0.0014*	0.0021*	0.0007*	0.0006*	0.0007*	0.0010*
15Scenes	0.0007*	0.1635	0.0007*	0.0010*	0.0007*	0.0007*	0.0007*
BMW	0.0355*	0.6374	0.0005*	0.9773	0.6694	0.0008*	0.0230*

C. Comparison with State-of-the-art Methods

Table I shows the mean value and standard deviation of classification accuracy over 15 rounds of experiments and Figure 3 depicts the boxplots of the classification accuracy.

As shown in Table I, all the existing methods cannot consistently outperform the baseline. Methods that target at learning discriminative features, *i.e.* Center, LGM and LMCL, are inferior to the baseline on the LM dataset, which may indicate that they sacrifice classification accuracy for small intra-class distance or large inter-class margin and underfit the data. Moreover, LGM performs poorly on the BMW dataset, which may be due to the mismatch between the distributional assumption of LGM and the real data. The dual loss is proposed to alleviate the vanishing gradient problem from using the cross entropy loss. Therefore, unless the problem occurs, we would expect its performance to be similar to the baseline. Such a pattern is found in our experiments, where only on the BMW dataset Dual improves the baseline by a large amount. Regarding the ensemble methods, we observe that the performance of Dropout is almost identical to that of the baseline and Snapshot performs worse than the baseline on three datasets. A potential justification for the unsatisfactory performance of Snapshot is as follows. For fairness of comparison, the learning rate is set to be the same for all methods and this value may be too small for Snapshot to reach local minima. Larger learning rates have been tested and we observe a degradation in the performance of Baseline. The proposed ReMarNet always outperforms the baseline; compared with state-of-the-art methods, it also achieves the highest accuracy on three datasets.

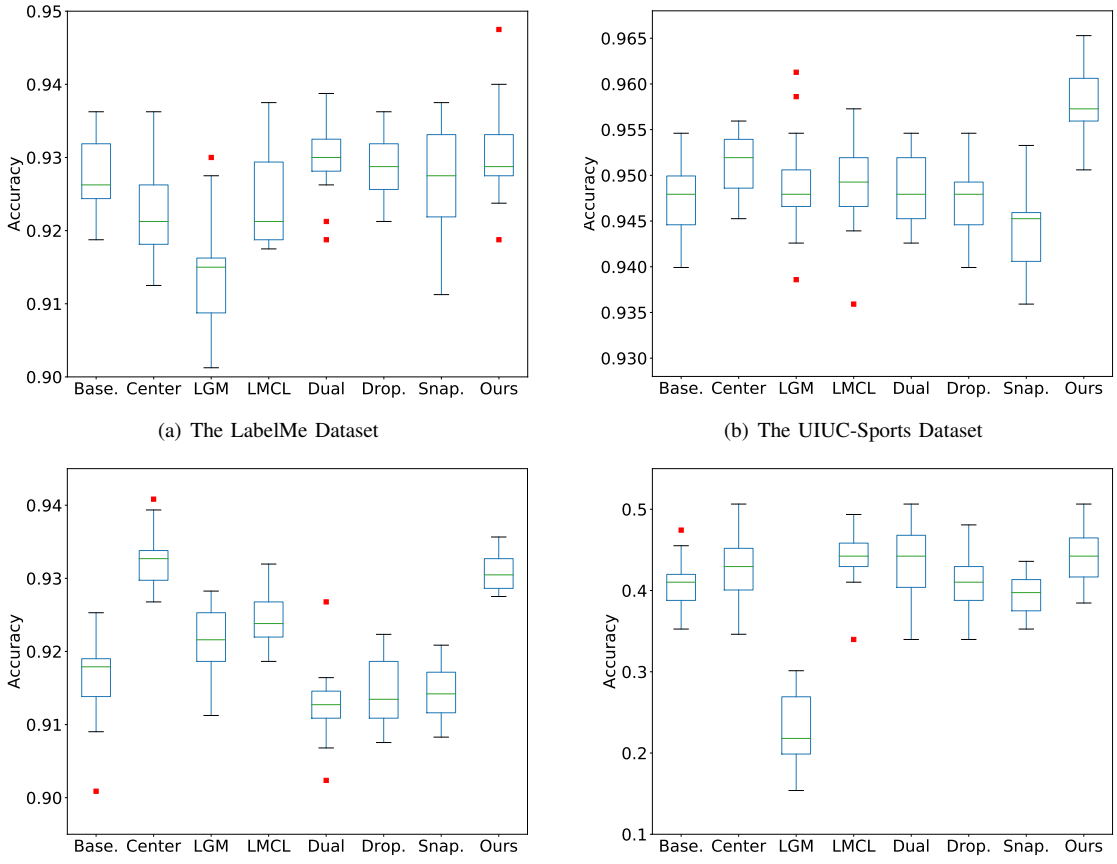
To further demonstrate the superiority of the proposed method, we conduct the Wilcoxon signed-rank tests [43] between the ReMarNet and other referred methods. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test, used to check the existence of significant difference between each pair of methods. Table II suggests that, at the significance level of 5%, ReMarNet is significantly different from others in the majority of cases.

We now focus on the reliability of the proposed method. Figure 3 shows boxplots of classification accuracy. Again, we observe that the existing methods outperform the baseline on some datasets only and deteriorate on at least one dataset, whereas our proposed ReMarNet achieves higher median accuracy on all four datasets and maintains similar spread as indicated by the interquartile range (IQR). On the LabelMe and BMW datasets, our method obtains higher first, second and third quartiles compared with the baseline. Its advantage is more pronounced on the UIUC and 15Scenes datasets. On the UIUC dataset, all methods have similar IQR but our method has a much higher median value. On the 15Scenes dataset, the proposed method has a smaller IQR than Baseline and the worst performance of ours is still larger than the baseline. Except on the LabelMe dataset, our method does not produce any outlier.

Another aspect of reliability is the method's stability to random choices of prototypes. The class prototypes are currently selected in a random manner at the beginning of the network training, and are fixed until the evaluation procedure finishes. To monitor the performance change of ReMarNet on the LM dataset, we sample 9 different sets of prototype images and run the ReMarNet for 15 rounds on each set of prototype images. Among 9 sets of experiments, the highest mean accuracy is 0.9339 and the lowest one is 0.9303; the standard deviation is 0.0067 in both cases. This result shows that the proposed ReMarNet remains stable as the class prototypes change.

D. Performance Evaluation under Different Training Sizes

To further evaluate classification performance of all methods in the small-sample size setting, we reduce the training sample size in the LabelMe and UIUC-Sports datasets. For LableMe, the number of training samples per class is reduced by 20, 40, 60 and 80 from its original size; the reduced datasets are denoted as LM-20, LM-40, LM-60, LM-80, respectively. For



(a) The LabelMe Dataset

(b) The UIUC-Sports Dataset

(c) The 15Scenes Dataset

(d) The BMW Dataset

Fig. 3. Boxplots of classification accuracy of the proposed ReMarNet and state-of-the-art methods. 'Baseline', 'Dropout' and 'Snapshot' are abbreviated to 'Base.', 'Drop.', 'Snap.', respectively. Each method has been evaluated for 15 rounds, and the distributions of accuracies are shown via boxplots. In each boxplot, the central mark is the median; the edges of the box are the 25th and 75th percentiles, respectively; and the outliers are marked in red individually.

TABLE III

COMPARISON OF CLASSIFICATION ACCURACY UNDER DIFFERENT TRAINING SAMPLE SIZES. THE NOTATION 'DATASETNAME-n' DENOTES THAT THE NUMBER OF TRAINING SAMPLES PER CLASS IS REDUCED BY n FROM ITS ORIGINAL SIZE.

Datasets	**Measure**	Baseline	Center	LGM	LMCL	Dual	Dropout	Snapshot	**Ours**
LM-20	**Mean**	0.9248	0.9116	0.9015	0.9065	0.9252	0.9247	0.9168	**0.9262**
	Std.	0.0059	0.0089	0.0095	0.0081	0.0045	0.0062	0.0101	**0.0054**
LM-40	**Mean**	0.9148	0.9180	0.9026	0.9138	0.9151	0.9148	0.9123	**0.9215**
	Std.	0.0055	0.0079	0.0087	0.0090	0.0062	0.0048	0.0059	**0.0064**
LM-60	**Mean**	0.9035	0.8947	0.8813	0.8971	0.9064	0.9036	0.9028	**0.9082**
	Std.	0.0056	0.0085	0.0126	0.0077	0.0067	0.0053	0.0085	**0.0081**
LM-80	**Mean**	0.8928	0.8933	0.8896	0.9011	0.8913	0.8917	0.8870	**0.9015**
	Std.	0.0098	0.0081	0.0081	0.0116	0.0119	0.0100	0.0170	**0.0084**
UIUC-10	**Mean**	0.9438	0.9531	0.9475	0.9463	0.9447	0.9443	0.9426	**0.9566**
	Std.	0.0041	0.0032	0.0062	0.0042	0.0064	0.0043	0.0050	**0.0038**
UIUC-20	**Mean**	0.9421	0.9456	0.9414	0.9429	0.9401	0.9401	0.9422	**0.9510**
	Std.	0.0040	0.0035	0.0081	0.0045	0.0051	0.0048	0.0060	**0.0042**
UIUC-30	**Mean**	0.9379	0.9340	0.9364	0.9366	0.9344	0.9372	0.9336	**0.9448**
	Std.	0.0048	0.0037	0.0054	0.0059	0.0046	0.0052	0.0060	**0.0053**
UIUC-40	**Mean**	0.9211	0.9251	0.9263	0.9266	0.9203	0.9213	0.9211	**0.9301**
	Std.	0.0081	0.0046	0.0064	0.0073	0.0061	0.0084	0.0073	**0.0060**

UIUC-Sports, we reduce 10, 20, 30, and 40 samples from each class and datasets are denoted in a similar way. The number of samples in the validation and test sets remains unchanged. The mean value and standard deviation of classification accuracy are listed in Table III.

As the training set gets smaller, it becomes more difficult to learn discriminative features. As anticipated, the accuracy of each method decreases with the number of reduced samples. On the LM dataset, Dual and Dropout, which originally outperform Baseline without data reduction, lose their advantages when the number of training samples is reduced by 80. Similar observations are found on the UIUC dataset, where Center, LGM, LMCL and Dual all perform worse than Baseline when reducing the training size by 30. In contrast, the proposed ReMarNet maintains its superiority over all methods across different sample sizes.

E. Ablation Study on the Impact of Different Backbone Networks

In the above experiments, ReMarNet and other compared methods adopt VGG16 as the backbone network. To further explore the potential of the proposed method, we use AlexNet [44] and DenseNet-121 [45] to construct ReMarNet and other compared methods, and run the experiment for 15 rounds on the LM dataset. The classification results are listed in Table IV. The performance of LGM is not listed when DenseNet-121 is used as the feature extractor as the method cannot fit the training data within 50 epochs.

From the table, we observe that all methods based on VGG16 perform better than their counterparts based on AlexNet and DenseNet-121; our method again outperforms all compared methods. This result shows that the proposed ReMarNet is effective even when the backbone network is changed.

F. Ablation Study on the Effectiveness of Two-branch Network

We now examine the structure of the proposed ReMarNet. Specifically, we compare the results of using only the relation module branch in the classification module (Base.-RM), using only the fully connected network branch (Base.-FC), and the proposed ReMarNet (Ours) which can be regarded as an ensemble of the RM and the FC network. The experimental results are shown in Table V.

First, we notice that the classification accuracy declines more rapidly in Base.-RM than Base.-FC. The reason might be that Base.-RM is more sensitive to the decrease in the number of pairs of feature embedding than Base.-FC. Second, combining these two networks, i.e. using the proposed ReMarNet, certainly provides a performance boost. The reason is that the RM branch and the FC branch have different classification paradigms, and put different assumptions during the training of network. As the ReMarNet trains the two branches conjointly, it will force the network to consider two kinds of assumptions in training and test procedures. Therefore, the generalization performance is enhanced in ReMarNet.

G. Ablation Study on the Effectiveness of Simultaneous Training and Prediction of Two-branch Network

As mentioned in Sec. I, our motivations behind the two-branch network are that the features allowing for two paradigms of classification should be more discriminative and that the decisions building on two paradigms should be more reliable. To verify if the proposed ReMarNet could achieve the above two goals, we design this experiment to evaluate the performance of the relation module branch trained on its own (Single-RM), the fully connected network branch trained on its own (Singe-FC), RM trained jointly with FC (Ours-RM), FC trained jointly with RM (Ours-FC), and the proposed ReMarNet (Ours). For Single-RM (Single-FC, resp.), we train the classification module with RM (FC network, resp.) only and the class label is predicted based on the relation score (probability vector, resp.); for Ours-RM (Ours-FC, resp.), we train the classification module with both RM and the FC network by using the proposed ReMarNet and then make predictions separately from the relation score (probability vector, resp.); for Ours, RM and the FC network are trained simultaneously and the prediction is made by summing up the two outputs. Figure 4 shows mean accuracy on the LabelMe and UIUC-Sports datasets.

By comparing single RM and Ours-RM, it is clearly evident that simultaneous training of two networks could generate more discriminative features, and, consequently, improve classification of each network. Similar observation can be found for the FC network, in particular on the UIUC dataset. By comparing Ours against Ours-RM and Ours-FC, we can see that the proposed ensembling network further enhances classification on the LabelMe dataset and the performance gain is more significant when the training set gets smaller. On the UIUC dataset, such improvement can still be observed in most cases. These encouraging results validate our motivation for simultaneous training and prediction on two networks.

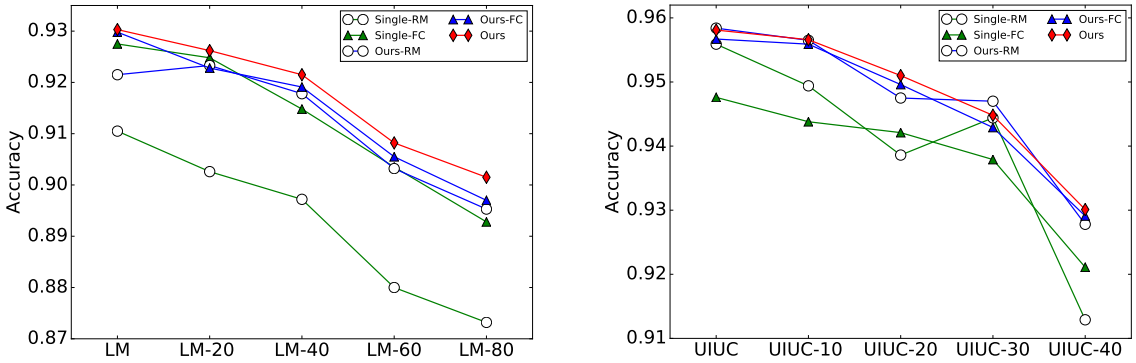
H. Performance Evaluation with Different Weights of Losses

In all the above experiments, our method is optimized by minimizing the total loss as presented in Eq. (4) for simplicity. A more deliberate choice is to compute the weighted sum of RM loss and FC loss. In this section, we change the weight of these two losses and monitor the performance of our method. Specifically, let a and b denote the coefficient of the RM loss and the FC loss, respectively; the previous setting corresponds to $a = 1$ and $b = 1$. We now fix $a = 1$ and vary b from 0 to 1, and likewise fix $b = 1$ and vary a from 0 to 1. The classification performance of our method on the LM dataset is listed in Table VI.

The accuracy remains competitive unless $b = 0$ where the FC network is no longer fine-tuned, suggesting the simultaneous training of two networks. Moreover, the classification performance can be improved by varying a or b. However, given the performance gain is relatively small and our task of interest is the small-sample classification, it would be preferable to avoid the introduction of the weight parameter and keep the model selection simple.

TABLE IV
COMPARISON OF CLASSIFICATION ACCURACY ON THE LM DATASET WITH DIFFERENT NETWORKS AS THE FEATURE EXTRACTOR.

Network	Measure	Baseline	Center	LGM	LMCL	Dual	Dropout	Snapshot	Ours
VGG16	**Mean**	0.9275	0.9219	0.9136	0.9207	0.9298	0.9288	0.9271	**0.9303**
	Std.	0.0047	0.006	0.0075	0.0155	0.0051	0.0045	0.0076	**0.0067**
AlexNet	**Mean**	0.8982	0.9013	0.8844	0.9006	0.8958	0.8976	0.9006	**0.9103**
	Std.	0.0051	0.0071	0.0172	0.0085	0.0058	0.0053	0.0058	**0.0050**
DenseNet-121	**Mean**	0.8846	0.8897	–	0.8880	0.8901	0.8846	0.8895	**0.8937**
	Std.	0.0109	0.0068	–	0.0090	0.0088	0.0100	0.0105	**0.0091**



(a) The LabelMe Dataset

(b) The UIUC-Sports Dataset

Fig. 4. Comparison of classification accuracy obtained by Single-RM, Single-FC, RM Branch (Ours-RM) and FC Branch (Ours-FC) in our method, as well as the proposed ReMarNet (Ours). In Single-RM and Single-FC, training and prediction are based on RM or FC only; in Ours-RM and Ours-FC, training is based on both RM and FC and prediction is based on RM or FC only; in Ours, training and prediction are based on both RM and FC.

I. Feature Visualization

We take the UIUC-Sport dataset as an example and visualize the features corresponding to different classes. We compare the features after training Baseline, Center, LGM, LMCL, Dual and the proposed ReMarNet. t-SNE [46] is used to depict the feature embedding f_φ in two dimensions, and results are given in Figure 5.

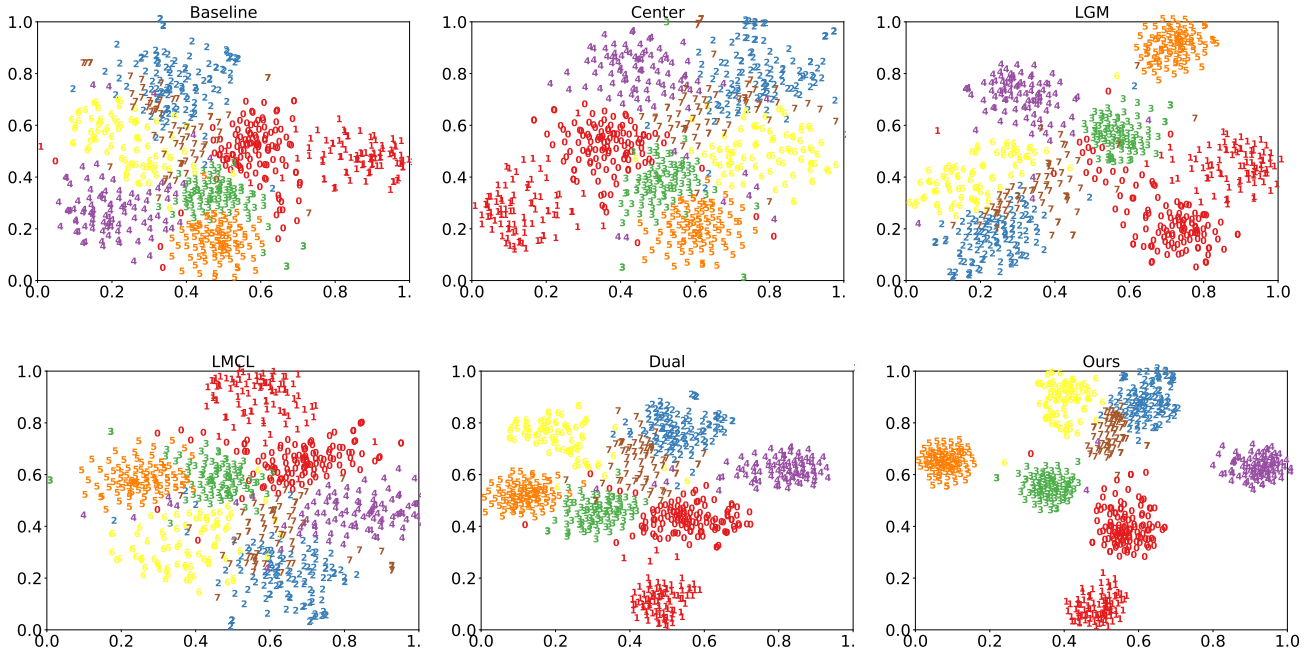
Figure 5 (a) clearly shows that, on the training dataset, feature embeddings learned from ReMarNet are more compact for samples of the same class and more separated between samples of different classes, compared with other methods. Such a clear pattern can be seen on the test set as well, supporting the superior performance presented in Table I. While not presented, we observe similar patterns on other datasets as well. This confirms that our method is capable of learning discriminative features.

J. Discussions

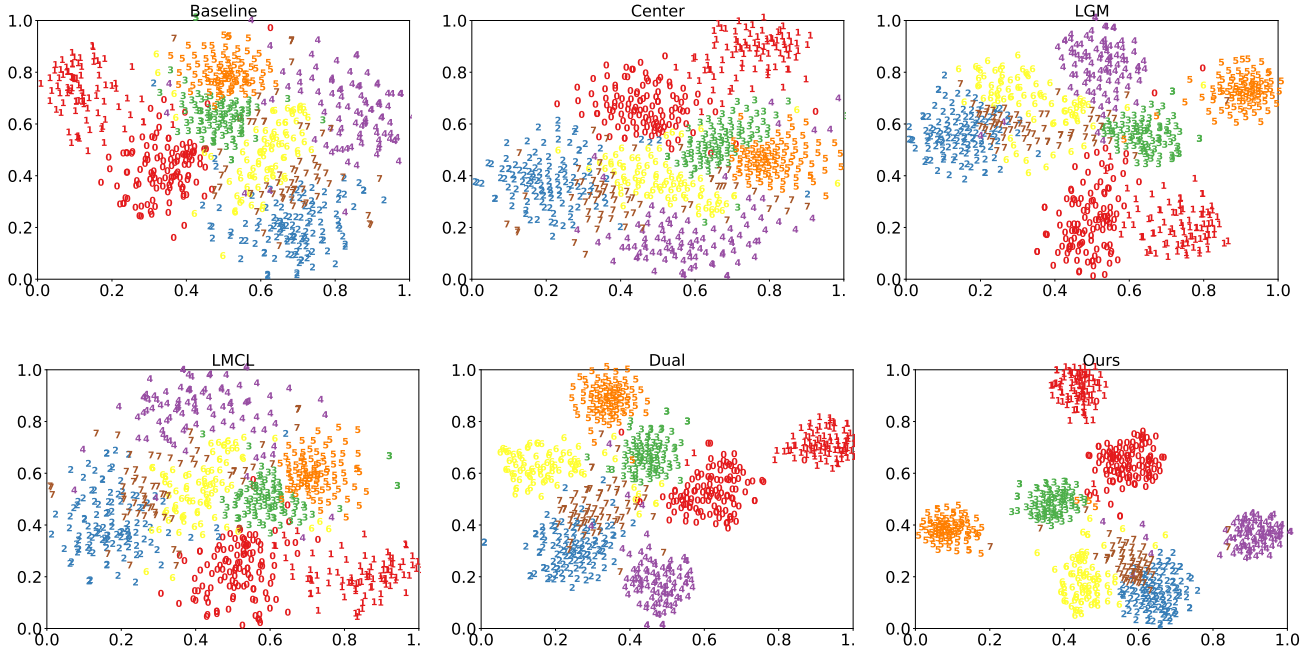
Our experiments demonstrate that learning discriminative features on a small training set is indeed a challenging task for existing state-of-the-art methods. They cannot guarantee to improve classification accuracy over the baseline network trained with the cross entropy loss on all the four datasets, and their performance deteriorates more when the training set size gets further reduced. More specifically, Center and LMCL involve additional hyperparameters during training and the selected values may not generalize well to the test data. On top of this issue, LGM assumes a Gaussian mixture distribution on features, which may not always fit for the data. To generate a good performance from Snapshot, it is essential that the local minimum can be obtained within the given number of epochs and the model can escape the local minimum when restarting the optimization. In the setting of small sample size, these two requirements slightly contradict each other since a small learning rate is needed to fine-tune features produced from the feature extractor module and a large learning rate is needed to escape from the local minimum.

Benefiting from a stricter requirement that the features should support a combined decision-making mechanism of two different classification paradigms, the proposed ReMarNet is capable of learning discriminative features and greatly enhances the baseline method on all of the evaluation datasets in this study. Its performance is also very competitive against state-of-the-art methods.

While our method achieves better performance, it has slightly more learnable parameters than the baseline and other compared methods. When the backbone network is VGG16, all compared methods have 15.5M parameters and our network has 16.1M parameters. When the backbone network is set as AlexNet or Densenet-121, the numbers of parameters in



(a) t-SNE feature visualization on the training set



(b) t-SNE feature visualization on the test set

Fig. 5. Visualization of feature embeddings on the UIUC dataset.

all compared methods are 2.8M and 3.1M, respectively; the numbers of parameters in the ReMarNet are 8.6M and 9.8M, respectively.

V. Conclusions, Limitations and Future Work

In this paper, we propose a new deep neural network for small-sample image classification called *Relation-and-Margin learning neural Network* (ReMarNet). It learns the discriminative features that can support both the classification paradigms based on the decision boundary and the similarity to class prototypes. Experimental results on four small datasets over a wide range of training sizes verify the efficacy of the proposed ReMarNet.

Here we would like to share two ideas of future work. Firstly, the class prototypes are currently selected from the samples, and it would be more effective to learn more representative ones. Secondly, although the ReMarNet is proposed for image classification, its framework is quite generic; therefore, applying it to other data types such as text data would be another valuable future work.

TABLE V
COMPARISON OF CLASSIFICATION ACCURACY OBTAINED FROM SINGLE-BRANCH RELATION NETWORK (BASE.-RM), SINGLE-BRANCH FULLY CONNECTED NETWORK (BASE.-FC) AND THE PROPOSED REMARNET (OURS). 'DATASETNAME-n' DENOTES THAT THE NUMBER OF TRAINING SAMPLES PER CLASS IS REDUCED BY n.

Dataset	Measure	Base.-RM	Base.-FC	Ours
LM	**Mean**	0.9105	0.9275	**0.9303**
	Std.	0.0088	0.0047	**0.0067**
LM-20	**Mean**	0.9026	0.9248	**0.9262**
	Std.	0.0119	0.0059	**0.0054**
LM-40	**Mean**	0.8972	0.9148	**0.9215**
	Std.	0.0079	0.0055	**0.0064**
LM-60	**Mean**	0.8800	0.9035	**0.9082**
	Std.	0.0120	0.0056	**0.0081**
LM-80	**Mean**	0.8732	0.8928	**0.9015**
	Std.	0.0135	0.0098	**0.0084**
UIUC	**Mean**	0.9559	0.9476	**0.9581**
	Std.	0.0055	0.0045	**0.0038**
UIUC-10	**Mean**	0.9494	0.9438	**0.9566**
	Std.	0.0051	0.0041	**0.0038**
UIUC-20	**Mean**	0.9386	0.9421	**0.9510**
	Std.	0.0077	0.0040	**0.0042**
UIUC-30	**Mean**	0.9444	0.9379	**0.9448**
	Std.	0.0087	0.0048	**0.0053**
UIUC-40	**Mean**	0.9129	0.9211	**0.9301**
	Std.	0.0081	0.0081	**0.0060**

TABLE VI
CLASSIFICATION ACCURACY OF REMARNET WITH DIFFERENT WEIGHTS ASSIGNED TO THE TWO-BRANCH LOSSES IN EQ. (4). a DENOTES THE WEIGHT OF THE RM LOSS AND b DENOTES THE WEIGHT OF THE FC LOSS.

Dataset	a	b	Mean	Std.
LM	**0**	**1**	0.9306	0.0053
	0.2		0.9305	0.0053
	0.4		0.9332	0.0058
	0.6		0.9322	0.0052
	0.8		0.9322	0.0050
	1		0.9303	0.0067
	1	**0**	0.9160	0.0046
		0.2	0.9345	0.0043
		0.4	0.9337	0.0061
		0.6	0.9325	0.0067
		0.8	0.9334	0.0050
		1	0.9303	0.0067

REFERENCES

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[2] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention lstm networks," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.

[3] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4683–4695, 2020.

[4] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2018.

[5] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 911–923, 2018.

[6] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *CVPR*, 2019, pp. 1328–1338.

[7] S. Jin, W. Liu, W. Ouyang, and C. Qian, "Multi-person articulated tracking with spatial and temporal embeddings," in *CVPR*, 2019, pp. 5664–5673.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[9] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 7167–7176.

[11] A. Rozantsev, M. Salzmann, and P. Fua, "Residual parameter transfer for deep domain adaptation," in *CVPR*, 2018, pp. 4339–4348.

[12] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *ICML*, 2013, pp. 1058–1066.

[13] X. Li, D. Chang, Z. Ma, Z. Tan, J. Xue, J. Cao, J. Yu, and J. Guo, "Oslnet: Deep small-sample classification with an orthogonal softmax layer," *IEEE Transactions on Image Processing*, 2020.

[14] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv preprint arXiv:1704.00109*, 2017.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[16] Z. Ma, J. Xie, Y. Lai, J. Taghia, J. Xue, and J. Guo, "Insights into multiple/single lower bound approximation for extended variational inference in non-gaussian structured data modeling," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2019.

[17] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018, pp. 1199–1208.

[18] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.

[19] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *International Conference on Learning Representations*, 2018.

[20] Y. Wang and Q. Yao, "Few-shot learning: A survey," *arXiv preprint arXiv:1904.05046*, 2019.

[21] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *ICCV*, 2017, pp. 3018–3027.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[23] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016, pp. 1249–1258.

[24] Y. Wei, F. Yang, and M. J. Wainwright, "Early stopping for kernel boosting algorithms: A general analysis with localized complexities," in *Advances in Neural Information Processing Systems*, 2017, pp. 6065–6075.

[25] J. Surh, H.-G. Jeon, Y. Park, S. Im, H. Ha, and I. So Kweon, "Noise robust depth from focus using a ring difference filter," in *CVPR*, 2017, pp. 6328–6337.

[26] G. Peng and S. Wang, "Weakly supervised facial action unit recognition through adversarial training," in *CVPR*, 2018, pp. 2188–2196.

[27] Z. Ren and Y. Jae Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in *CVPR*, 2018, pp. 762–771.

[28] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[29] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *ICML*, 2016, pp. 507–516.

[30] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *CVPR*, 2017, pp. 212–220.

[31] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *CVPR*, 2018, pp. 5265–5274.

[32] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*. Springer, 2016, pp. 499–515.

[33] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," *arXiv preprint arXiv:1710.00870*, 2017.

[34] W. Wan, Y. Zhong, T. Li, and J. Chen, "Rethinking feature distribution for loss functions in image classification," in *CVPR*, 2018, pp. 9117–9126.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[37] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *CVPR*, vol. 2. IEEE, 2005, pp. 524–531.

[38] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, vol. 2. IEEE, 2006, pp. 2169–2178.

[39] J. Krause, M. Stark, J. Deng, and F.-F. Li, "3D object representations for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.

[40] X. Li, L. Yu, D. Chang, Z. Ma, and J. Cao, "Dual cross-entropy loss for small-sample fine-grained vehicle classification," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4204–4212, 2019.

[41] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

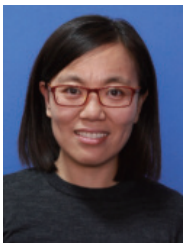
[42] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[43] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[46] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.



Xiaoxu Li received her Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 2012. She is an associate professor in the School of Computer and Communication at Lanzhou University of Technology. Her research interests include machine learning fundamentals with a focus on applications in image and video understanding.

Liyun Yu received his B.E. degree in computer science and technology from Lanzhou University of Technology, China, in 2016. He is currently a graduate student in Lanzhou University of Technology. His research interests include machine learning and small-sample image understanding.

Xiaochen Yang received the B.Sc. degree in Actuarial Science from London School of Economics in 2013 and the M.A.St. degree in Mathematical Statistics from the University of Cambridge in 2015. She is currently pursuing the Ph.D. degree in statistical science at University College London. Her research interests include metric learning, statistical classification, and hyperspectral image analysis.

Zhanyu Ma has been an Associate Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He is also an adjunct Associate Professor at Aalborg University, Aalborg, Denmark, since 2015. He received his Ph.D. degree in Electrical Engineering from KTH (Royal Institute of Technology), Sweden, in 2011. From 2012 to 2013, he has been a Postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics.

Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is an associate professor in the Department of Statistical Science at University College London. His current research interests include statistical classification, high-dimensional data analysis, computer vision, and pattern recognition.

Jie Cao received her M.E. degree from Xi'an Jiaotong University, China, in 1994. She is a professor and a vice president of Lanzhou University of Technology. Her research interests include machine learning, pattern recognition, speech and speaker recognition, information fusion and computer vision.

Jun Guo received the B.E. and M.E. degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, and the Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. He is a professor and a vice president of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and bioinformatics. He has published over 200 papers on the journals and conferences including SCIENCE, Nature Scientific Reports, IEEE Trans. on PAMI, Pattern Recognition, AAAI, CVPR, ICCV, SIGIR, etc.