IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received February 26, 2021, accepted March 11, 2021, date of publication March 22, 2021, date of current version April 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068042

Normalized Maximal Margin Loss for Open-Set Image Classification

CHUNMEI MA, HUAZHI SUN, JINQI ZHU, LONG ZHANG, BOJUE WANG, DONGHAO WU, AND JINGWEI SUN
School of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China

Corresponding author: Jinqi Zhu (zhujinqi1016@163.com) and Huazhi Sun (sunhuazhi@tjnu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61902282; in part by the Natural Science Foundation of Tianjin under Grant 18JCYBJC85900, Grant 18JCQNJC70200, and Grant 20JCZDJC00400; in part by the Tianjin Municipal Education Commission Project for Scientific Research Plan under Grant 2018KJ155; in part by the Science and Technology Development Fund of Tianjin under Grant JW1702; and in part by the Science and Technology Project of Guangdong Province under Grant 2017KQNCX194.

ABSTRACT This work aims to address the image classification problem under open-set protocol: classes in test set do not appear in the training set. Intuitively, convolutional Neural Network (CNN) with softmax loss is a straight-forward solution. However, the unknown class (is not predefined in the training set) makes the boundaries of intra-class and inter-class more blurred, which brings more challenges for image classification. Although some softmax variants, such as center loss, CosFace loss etc., focus on learning discriminative features by minimizing the intra-class distance, they do not explicitly maximize inter-class distance, which is more important for open-set problem verified by our experiments. Besides, even though deep metric learning, such as with the contrastive loss and the triplet loss, can learn discriminative features of intra-class and inter-class, it needs a time-consuming image sampling process during training. In this paper, we propose a novel normalized maximal margin (NMM) loss for open-set image classification, which not only *explicitly* minimizes intra-class distance and maximizes inter-class distance, but also defines their margins. Specially, after analyzing the advantage of angular space that the softmax loss normalized by the feature and weights through geometric interpretation, we make NMM work in angular space. Then, the validity of NMM for discriminative features learning is demonstrated from the view of geometric interpretation as well. After that, we innovatively determine the upper bound of inter-class margin by theoretical analysis. Finally, extensive experiments are conducted on popular datasets: CIFAR-100 (object recognition), ImageNet (image classification), LFW (face recognition) and MSMT17 (person re-identification) to verify the effectiveness of NMM. The experimental results show that NMM achieves very competitive performance.

INDEX TERMS Deep metric learning, open-set image classification, inter-class distance, convolutional neural network.

I. INTRODUCTION

Image classification is a fundamental task in computer vision and pattern recognition. It can easily be categorized as ‘‘closed-set’’ and ‘‘open-set’’ settings [1]. For ‘‘closed-set’’ protocol, all classes in the test set are predefined in the training set. In comparison, classes of the test set are disjoint from those of the training set under ‘‘open-set’’ protocol [2], [3], which is more in line with the real application scenarios. Take the person re-identification as an example, as shown in the box 1 in Fig.1. A model is trained by training set, then it is applied to the real world where samples (in the testing phase) are not predefined in training set. At this time, the test set is divided into two parts: the probe and the gallery. Under open-set protocol, the model performs image verification between the probe image and every identity in the gallery. There are similar situations in the applications of face recognition and clothing searching etc. (see Fig.1). In this paper, we focus on ‘open-set’ protocol, which requires the classifier in the test phrase to classify the unknown class (is not predefined in the training set) rather than to misclassify as the known class. Clearly, ‘open-set’ protocol is more challenging than the ‘close-set’ one because the unknown class makes the boundaries of intra-class and inter-class more

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Zhang.

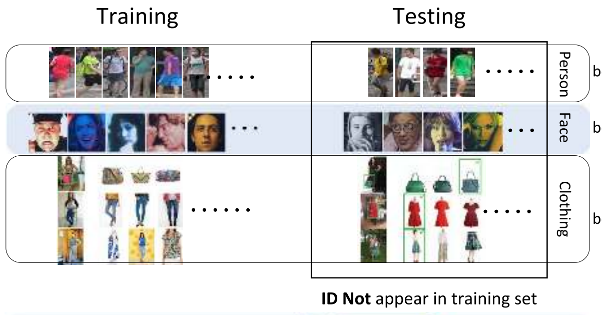
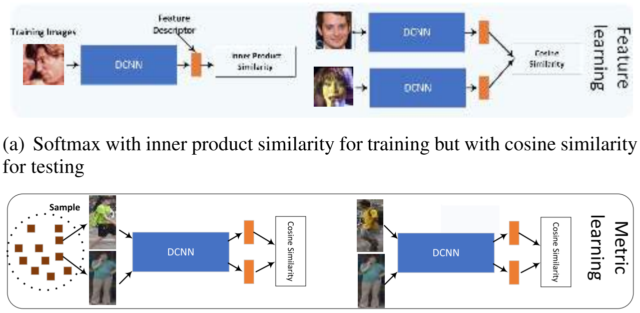


FIGURE 1. **Open-set image classification. Three open-set applications: person re-identification (box 1), face recognition (box 2) and clothing search (box 3).**



(a) Softmax with inner product similarity for training but with cosine similarity for testing

(b) Traditional metric learning with sampling

FIGURE 2. **The methods of discriminative feature learning for open-set image classification. (a) softmax loss-based; (b) metric learning.**

blurred. Therefore, we have to learn a *discriminative enough* space to project testing images.

Thanks to the development of Convolutional neural network (CNN), it has already surpassed human-level performance on several open-set image classification benchmarks [4], [5]. Pioneering work [6], which focuses on the open-set problem, is to learn the identity features via the softmax loss (Following the [7], we define the softmax loss as the combination of the last fully connected layer, softmax function and cross entropy loss). However, even if softmax loss is the most widely used loss for open-set problem, it only guides CNNs to learn separable features that are not discriminative enough because it has a gap of distance between training and testing, reducing the classification performance. Taking face recognition for example, people usually use softmax (inner product distance) to train a CNN, however, cosine distance is used during test. This distance gap is illustrated in Fig.2 (a).

To address the above problem, some works propose to normalize the learned features to bridge the distance gap. Specially, these methods can be simply divided into two classes: (i) normalizing feature only [8] and (ii) normalizing both features and the weights of classifier [9], [10]. In this way, these methods work in the angular space rather than Euclidean space. From the view of geometric interpretation, we have demonstrated the features in the angular space are more discriminative than those in Euclidean space (see section III-B), which is consistent with the conclusion of [10]. However, they do not explicitly encourage intra-class compactness and inter-class separability. To this end, some softmax variants aimed at minimizing the intra-class distance, such as center loss [11] and CosFace loss [12] are proposed. But they do not specify maximize inter-class distance.

Apart from softmax-based methods for open-set problem, metric learning, which aims to learn a similarity function, is also widely used. Recently, prevailing deep metric learning usually uses neural networks to automatically learn discriminative features x_1 and x_2 of two samples. Then, a simple distance metric such as Euclidean distance or cosine distance is applied to x_1 and x_2 to determine the similarity of the two samples. Most widely used loss functions for deep metric learning are contrastive loss [13] and triplet loss [14]. Some methods combine softmax loss with contrastive loss or triplet loss to enhance the discrimination power of features. They are very common for face recognition and person re-identification, achieving promising recognition performance. Unlike softmax, metric learning intrinsically does not have the distance gap. However, both contrastive loss and triplet loss cannot constrain on each individual sample, and thus require carefully designed pair/triplet mining procedure, as shown in Fig.2 (b), which is time-consuming and performance-sensitive.

In this paper, we propose a new normalized max margin (NMM) loss. Following the existing work [15], we normalize the features and the weights of a classifier to make NMM work in the angular space, bridging the distance gap between training and testing. Unlike [3], [12], [16], [17] which only explicitly minimize the intra-class distance, NMM explicitly minimizes the intra-class distance and maximizes the inter-class distance which is more important for open-set image classification tested by experiments and defines their margins. Furthermore, the upper bound of inter-class margin is determined by theoretical analysis. In addition, NMM is actually one metric learning method. Unlike contrastive loss and triplet loss, NMM intrinsically does not have the sample mining procedure, making the training procedure more efficient. Apart from these, NMM also has clear geometric interpretations. Supervised by NMM loss, the learned features can construct a discriminative angular distance metric that is equivalent to geodesic distance on a hypersphere manifold. NMM loss can be interpreted as constraining learned features to be discriminative on a hypersphere manifold.

The original contributions that we have made in the paper are highlighted as follows:

- We propose a normalized max margin (NMM) loss for open-set problem. NMM works in a discriminative angular space and *explicitly* minimizes the intra-class distance and maximizes the inter-class distance, which is more important for open-set problem tested by experiments. Unlike most existing metric learning methods, NMM does not have the sample mining procedure. In addition, NMM has clear geometric interpretations to make the proposed model more interpretable.

- We innovatively determine the upper bound of inter-class margin by theoretical analysis rather than experiment, which is more appropriate for the open-set problem.
- We conduct extensive experiments on four databases: CIFAR-100 (object recognition), ImageNet (image classification), Labeled Face in the Wild (LFW, face recognition) and Multi-Scene Multi-Time person ReID dataset (MSMT17, person re-identification). Experimental results verify the effectiveness of the proposed method.

II. RELATED WORK

Open-set image classification has been studied for very long time. How to learn a discriminative enough feature space is key to the classification performance. In deep learning era, the methods for the open-set problem can easily be classified into two categories: softmax loss-based and metric learning.

A. SOFTMAX LOSS-BASED

Deep learning models with different softmax loss functions can learn different discriminative features. In the early stage, softmax loss with dot product similarity is widely used for the open-set problem [6], [18]. However, there is a problem with different similarity calculation between training and testing, such as dot product similarity is used for training, cosine similarity is used for testing, reducing the classification performance. Therefore, feature normalization is proposed to learn more discriminative features for the open-set problem, e.g. L2-normalized Euclidean distance and cosine distance. Parde *et al.* [19] observe that the L2-norm of features learned using softmax loss is informative of the quality of the object (face). Features for frontal faces have a high L2-norm while blurry faces or extreme-pose faces have low L2-norm. Ranjan *et al.* [8] add the L2-constraint to the feature descriptors and restrict features to lie on a hypersphere of a fixed radius. Though these methods achieve promising performance on the open-set problem, they do not explicitly minimize intra-class distance and maximize inter-class distance.

To make the features of intra-class more compact and increase the separation of inter-class, some softmax variants are proposed. In [11], the authors propose the center loss which directly constrains the distance between sample features, making similar features close to their center point. However, it ignores the unevenness of the sample distribution, resulting in unclear boundaries between classes. Different from center loss, Wan *et al.* [20] propose the L-GM loss in which features are supposed to be a mixed Gaussian distribution. However, for the open-set problem, the distribution of the unknown class with multiple categories is completely different from the distribution of the predefined classes. After that, some losses with margins are proposed [3], [7], [12], [16], [17], [21]. Liu *et al.* [3], [7] propose a large margin Softmax (LMSoftmax) by adding multiplicative angular constraints to each identity to improve the feature discrimination on vision classification and face verification. Due to the non-monotonicity of the LMSoftmax, an operator is employed to guarantee the monotonicity which makes LMSoftmax hard to optimize. To overcome this problem, [12], [16], [17] propose to move the angular margin to cosine space, which makes the implementation and optimization much easier than LMSoftmax. Although these works achieve promising performance on the open-set problem, they only explicitly minimize intra-class distance. In [21], the circle loss explicitly minimizes intra-class distance and maximizes inter-class distance. However, it requires pair-wise features, one of which is the positive sample feature and the other one is the negative sample feature. Besides, for the above margin losses, they mainly focus on the margin of intra-class. However, we find through experiments that open-set problem is more sensitive to inter-class margin. Furthermore, the margin value is determined mainly by experiments, which may affect their applicability for real scenes with unknown class.

B. METRIC LEARNING

It is widely used for open-set problem due to its strong feature learning capacity. With the strong feature learning capacity of deep learning, deep metric learning can perform even better than the traditional methods [22], [23]. For this, a good distance metric is critical for its success. A comprehensive survey of the deep metric learning methods can be accessed in [24]. Recently, more complicated loss functions were proposed to train a better image representation. Most widely used loss functions for deep metric learning are contrastive loss [13] and triplet loss [14]. They both impose Euclidean margin to features and need sample mining mechanism which is time consuming. In comparison, our NMM loss works in a more discriminative angular space and does not need the sample mining module.

III. NORMALIZED MAXIMAL MARGIN (NMM) LOSS

In this section, we first briefly introduce softmax, feature and weight normalization. Then we give a deep analysis of feature normalization from the view of geometric interpretation. After that, we detail the proposed NMM loss. Finally, we make deeper analysis on NMM loss including geometric interpretation and bound discussion.

A. PRELIMINARIES

To better understand the proposed NMM loss, we first briefly review the original softmax and its variants. The original softmax is formulated as:

$$L_S = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{W_{y_i}^T f_i + b_{y_i}}}{\sum_{j=1}^{c} e^{W_j^T f_i + b_j}} \tag{1}$$

where f_i is the deep feature of sample i (the i-th output of the fully connected layer), y_i is the true label of sample i, $W_{y_i}^T$ is the weight of f_i, the $W_{y_i}^T f_i$ is called as the target logit [14] of the i-th sample, b_j is the bias of the j-th class. The softmax function can be generalized by setting the logit as a function $\phi(W_j, f)$:

$$L_S = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{\phi(W_{y_i},f_i)}}{\sum_{j=1}^{c}e^{\phi(W_j,f_i)}} \tag{2}$$

where $\phi(W_j, f_i) = W_j^T f_i + b_j$

1) WEIGHTS NORMALIZATION

To improve the discriminative capacity of the learned features for the open-set problem, the weights from the classifier are normalized. To facilitate the analysis on the loss function, we set the bias $b_j = 0$ as [3]. Then, the logit of Eq. (2) can be changed to:

$$\phi(W_j, f_i) = \frac{W_j}{\|W_j\|}f_i = \|f_i\|\cos < W_j, f_i > \tag{3}$$

We can see that the original softmax function with dot product similarity is converted to the one with cosine similarity (angular space), which is claimed to be more discriminative [3].

As shown in [3], they set $\|W_j\| = 1$ by l_2 normalization, making the predictions only depend on the angle between the feature vector and the weights.

2) FEATURE NORMALIZATION

Apart from weights normalization, feature normalization is also widely used to solve the open-set problem. The feature normalization can be formulated as:

$$\phi(W_j, f_i) = \gamma W_j^T\frac{f_i}{\|f_i\|} = \gamma\|W_j\|\cos < W_j, f_i > \tag{4}$$

where γ is a scalar to help network convergence which was discussed in [8]. It is observed in [19] that the l_2-norm of features learned is informative for the quality of the face. Specifically, features with good quality (e.g. frontal faces) have a higher l_2-norm than blurry faces with large poses. The advantages of feature normalization are also discussed in [8]–[10].

3) WEIGHTS AND FEATURE NORMALIZATION

The weights and features can also be $l2$ normalized together [15]:

$$\phi(W_j, f_i) = \gamma\frac{W_j^T f_i}{\|W_j\|\|f_i\|} = \gamma\cos < W_j, f_i > \tag{5}$$

where γ is a scalar, $l2$ normalization on features and weights lead to a so-called hypersphere metric learning. Normalizing the features and weights can remove the radial variations and push every feature to distribute on a hypersphere manifold.

4) ADDITIVE ANGULAR MARGIN

Apart from performing normalization in softmax loss, some recent researches add angular margin to softmax loss, e.g. SphereFace [3], CosFace [12] and ArcFace [16]. Take CosFace [12] for example, the formulation is:

$$L_{CosFace} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{\gamma(\cos(\theta_{y_i})-\delta)}}{e^{\gamma(\cos(\theta_{y_i})-m)} + \sum_{j=1,j\neq y_i}^{c}e^{\gamma\cos(\theta_j)}} \tag{6}$$

where $\cos(\theta)$ denotes the cosine similarity between the feature and the weight, δ is the margin in the angular space, γ is a scalar.

B. FEATURE NORMALIZATION ANALYSIS

The feature normalization [9] was originally proposed to bridge the gap between training and test. For example, the Euclidean distance was used for training, but cosine distance for test. The existence of this gap will potentially reduce the recognition performance. We will analyze the reasons with geometric interpretation as follows.

Fig.3(a) and Fig.3(b) represent training using dot product similarity (softmax) and testing using cosine similarity respectively. In Fig.3(a), four hyperplanes clearly separate four classes in training space, and the red arrows are the corresponding normal vectors. These hyperplanes do not always get through the origin of coordinates due to the existence of the bias of one classifier, e.g., b in Eq.(1). In Fig.3(b), the features are normalized to a sphere and the bias is removed during test, leading to class overlapping on the sphere (the normalized feature space). Therefore, if it keeps consistence during training and testing in the way of removing the bias and normalizing the features, we can obtain more discriminative features.

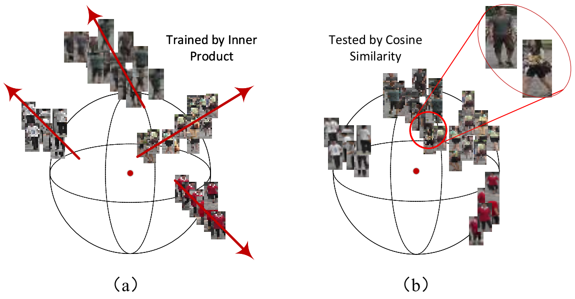
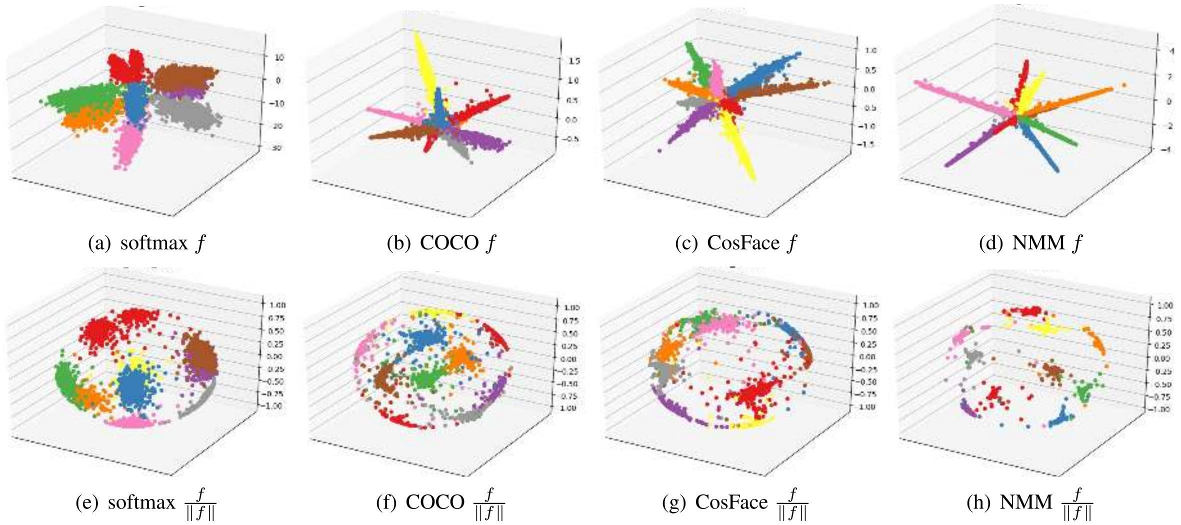


FIGURE 3. The distance gap between training and testing.

C. NORMALIZED MAXIMAL MARGIN LOSS

As analyzed above, the normalization of feature and/or weight in the softmax loss can change the inner product distance to angular distance, thereby improving the feature discrimination. However, most of these methods [12], [20] do not explicitly maximize inter-class distance. Metric learning explicitly minimizes the intra-class distance and maximizes the inter-class distance. In addition, metric learning does not have the distance gap during training and test. However, metric learning methods intrinsically need the sample mining mechanism, e.g. hard negative sampling, which is time-consuming and performance-sensitive.

In this paper, we propose a novel loss function named Normalized Max Margin loss (NMM) that explicitly minimizes the intra-class distance and maximizes the inter-class distance and defines their margins respectively. The challenges of constructing NMM are how to combine the distances of inter-class and intra-class and how to integrate inter-class



(a) softmax f (b) COCO f (c) CosFace f (d) NMM f

(e) softmax $\frac{f}{\|f\|}$ (f) COCO $\frac{f}{\|f\|}$ (g) CosFace $\frac{f}{\|f\|}$ (h) NMM $\frac{f}{\|f\|}$

FIGURE 4. Visualization of unnormalized features (Row 1) and normalized features (Row 2). Different colors indicate different classes. The features are trained using MNIST dataset.

margin with inter-class distance. For this purpose, NMM loss is represented as:

$$L_{NMM} = \frac{1}{n}\sum_{i=1}^{n}[\underbrace{\max\{0, l - \tilde{s}_{y_i,i}\}}_{intra-class} + \underbrace{\sum_{j\neq y_i}\max\{0, \tilde{s}_{j,i} - m\}]}_{inter-class} \tag{7}$$

where l and m are the margins of intra-class and inter-class terms respectively. The learned feature is given as:

$$\tilde{s}_{k,i} = \frac{W_k^T f_i}{\|W_k\|\|f_i\|} \tag{8}$$

where W_k is k-th column of the weight from last fully connected layer and f_i is the output of layer penultimate. If $k = y_i$ that is the true class label of sample i, $\tilde{s}_{y_i,i}$ denotes the intra-class distance of smaple i. If $k = j$ that is the class label different from the class of sample i, $\tilde{s}_{j,i}$ is the inter-class distance. It is worth noting that the bias is removed. Thus, our NMM loss works in angular space. In addition, NMM is a linear combination of $\tilde{s}_{k,i}$ which is differentiable [15]. Thus, our NMM is differentiable.

Clearly, NMM loss consists of two terms: the first item $\max\{0, l - \tilde{s}_{y_i,i}\}$ describes the intra-class cosine similarity of sample i, and restricts the distance $\tilde{s}_{y_i,i}$ with margin l. l=1 means intra-class loss will be activated when the angular between sample and its corresponding W larger than 0 degree. The second term $\sum_{j\neq y_i}\max\{0, \tilde{s}_{j,i}-m\}$ models the inter-class cosine similarity of sample i, and restricts the distance $\tilde{s}_{j,i}$ with margin m. m=0 means inter-class loss will be activated when the angular between sample and other w less than 90 degree. The two margins l and m encourage a CNN to learn a clear class bound. The combination of two terms encourages the intra-class compactness and inter-class separability.

Unlike aforementioned softmax variants and metric learning, our NMM has three key advantages: (1) it explicitly minimizes the intra-class distance and maximizes the inter-class distance; (2) it defines their margins respectively, as intra-class term in Eq. (7)) with margin l and inter-class term in Eq. (7) with margin m; (3)it intrinsically removes the sampling mechanism.

D. FEATURE VISUALIZATION ANALYSIS

In this section, we analyze the effects of our NMM by feature visualization. Feature and weight normalization make our NMM work in the angular space like SphereFace. For comparison, we visualize the feature distributions trained by several loss functions that are work in angular space in Fig.4: softmax loss, COCO loss [9], CosFace loss [12] and our NMM loss. We used MNIST (10 classes) as the training data, and a LeNet variant is used to output 3D features, for training. These networks are supervised by those four loss functions. We train our NMM loss with $l = 1.0$, $m = 0.5$. After training, we plot the obtained 3D features as shown in Fig.4 (Row 1), and then normalize these features on a hypersphere (Row 2).

From Fig.4 (Row 1), we can see the features of our NMM loss is more compact than others. In Row 2, all the features are normalized on a sphere. We can see that the features of NMM are more separable than other features. Thus, we can conclude that NMM can learn very discriminative features by explicitly optimizing intra- and inter-distance.

E. MARGINS ANALYSIS

To analyze the effects of NMM margins m and l setting on the training process, we first analyze the initial distribution of samples in the feature space. We find that the samples initially distribute on a very small cluster as shown in Fig.5(a). In Fig.5(b), if the features are fed to the RELU, then the negative parts are removed. If the features are normalized,

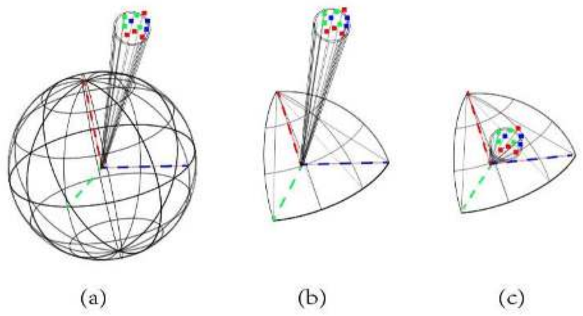


FIGURE 5. The initial distributions of samples. Dots indicate samples. (a) features before passing activation function; (b) features after RELU activation; (c) features are normalized.

then they distribute on a sphere as shown in Fig.5(c). In all these 3 cases, the samples all initially distribute on a small cluster.

To verify the assumption that samples initially distribute on a small cluster, we conduct a quantitative experiment. We chose a popular Person Re-Identification database Market-1501 [25], which is a typical open-set task. We randomly initialized ResNet50 model to map all train samples (11159 images) to the feature space. Then, we compute the cosine similarity of all the pairs (11159^2 in all). Fig.6 shows the statistical histogram of cosine similarity. We can see that the minimal cosine value is 0.994 which means the maximal angular distance between two features (1 pair) is 6 degrees. Therefore, all samples are gathered in a small cluster ($< 6°$).

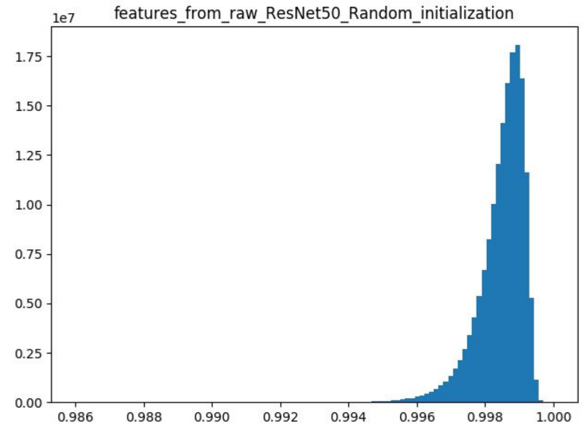


FIGURE 6. The statistical histogram of the cosine similarity. We use all the 11159 images from Market-1501 for visualize. The cosine similarity of all the pairs 11159^2 are computed. We use a randomly initialized ResNet50 for feature extraction.

Based on the aforementioned initial distribution, we then analyze how the margins affect the training. From our NMM loss in Eq. (7), the samples move under two forces from intra-class term and inter-class term, respectively. The former (intra-class) guides the sample to move to its center. The latter (inter-class) ensures the centers of all classes far from each other. The latter can be viewed as one regularizer. Clearly, the regularizer should be larger than 0, meaning that we have to keep the regularizer to avoid model overfitting. Thus, we think one critical point is that the force of inter-class term is 0 (no regularizer). In other words, the regularizer can avoid the model overfitting, but it will make the model hard to converge. The stronger the regularizer is, the harder the model converges.

Why does the regularizer make the model hard to converge? As shown in Fig.7, there are 3 classes (each color indicates one class) and 10 samples (4 green, 4 red and 3 blue points). Note that we only draw 3 classifier centers W (the same color of W and the samples are from the same class). Fig.7(a) is the initial distribution. In Fig.7(b), we add our NMM loss to the samples. After training, the effective feature space is compressed to a size equal to summation of inter-class and intra-class margin as shown in Fig.7(c). This is the reason why strengthening the regularizer can avoid the model overfitting and help to learn a more discriminative feature space. In Fig.7(b), the colored shadow regions that mean the inter-class force for those samples, which do not have the same color as shadow (not the same class), will be activated. The white regions mean that samples are guided by intra-class force only. The inter-class margin m controls the size of its corresponding region: the smaller the m, the larger the region. For example, if the inter-class region larger than the situation in Fig.7(b), it will cause some overlapped regions where the samples will be guided by more than two forces corresponding to the overlapped regions. The joint force will make a wrong direction to the sample. This is the reason why the regularizer makes the model hard to converge. Therefore, we can see that the inter-class margin plays more significant role for the open-set problem that owns unknown class with large distribution.

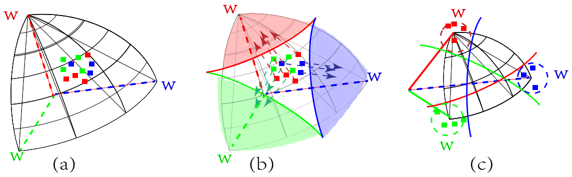


FIGURE 7. The critical state of margin m setting. Different colors of dots indicate different classes. (a) the initial distribution of samples before training; (b) the margin m is performed at the critical state where the shadow regions are tangent to each other; (c) the distribution of samples after training.

Since the critical state is that the inter-class term is equal to 0, which means there are no overlapped sector region such as Fig.8. When the inter-class margin is larger than the situation of Fig.8, the regularizer is larger than 0 and we can learn a discriminative feature space, but the model is hard to convergence. Instead, when the inter-class margin is smaller than the situation of Fig.8, the regularizer is 0 and the model is easy to train, but the discriminative feature space is weak. Thus, the critical state is the situation we can make the model easy to converge and learn a relatively discriminative feature space.

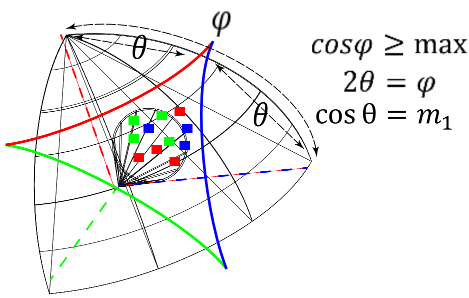


FIGURE 8. The geometric description of critical state that two shadow regions are tangent to each other.

Based on the aforementioned analysis and the critical state, we compute the upper bound of inter-class margin m. It means we do not need to consider the value of m larger than the upper bound. We have the normalized feature x, unit weight vectors W, the number of classes as C and K is the dimension of the learned features. Suppose that the learned features separately stay on the surface of a hypersphere and their centers are around the corresponding weight vector W, in which W evenly distributes in the sphere. Then we have

$$m \leq \sqrt{\frac{1}{2}} \tag{9}$$

where $K \geq 3$ and $2 < C < K + 1$.

Proof: For $K \geq 3$ and $||W|| = 1$, the inequality below holds [12]:

$$C(C-1)W_i^T W_j = \sum_{i,j,i\neq j} W_i^T W_j \tag{10}$$

$$\sum_{i,j,i\neq j} W_i^T W_j = (\sum_i W_i)^2 - (\sum_i W_i^2) \leq 0 \tag{11}$$

Combining Eqs. (10) and (11), we can obtain $W_i^T W_j \leq 0$. Fig.8 shows the critical state when every two shadow regions are tangent to each other. Thus, we have $2\theta = \varphi$, where θ is the angle corresponding to inter-class margin m and φ is the angle of two adjacent cluster centers W. Thus, we have

$$\cos\varphi = W_i^T W_j \tag{12}$$

Since $\cos\theta = m$ and $2\theta = \varphi$, we combine Eqs.(10), (11), (12) to get:

$$m \leq \sqrt{\frac{1}{2}} \tag{13}$$

The inter-class margin makes the learned features have better separability, reducing the intra-class distance at the same time. The intra-class margin makes its features more compact, reducing the number of features falling near to the inter-class margin to increase the accuracy of classification. But obviously, the open-set problem is more sensitive to the inter-class margin than the intra-class margin. The determination of intra-class margin mainly relied on experiments which will be introduced in section IV-A2.

IV. EXPERIMENTS

In this section, we conduct a series of experiments to verify the effectiveness of the proposed NMM. Specifically, we evaluate our method on a toy dataset (a subset of CIFAR-100 for object recognition), the ILSVR2012 dataset (a subset of ImageNet with more classes for image classification) and two 'real' datasets (LFW [26] for face recognition, MSMT17 [27] for Person Re-identification). For the experiments on CIFAR-100, we also evaluate the effects of the hyper-parameters (margins l and m) of NMM. Then, we use the best hyper-parameters from CIFAR-100 for the following experiments on ILSVR2012, LFW and MSMT17. In this experiment, we compare our method with the state-of-the-art competitors: the softamx+cross-entropy loss (baseline), Center loss [11], COCO loss [9], SphereFace loss [3], CosFace loss [12], ArcFace loss [16] and other state-of-the-art methods on face recognition and Person Re-identification task.

A. CIFAR-100 (OBJECT RECOGNITION)

1) DATASET AND SETTINGS

The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs). To simulate the open-set problem, we randomly divide the 100 classes into 90 (train) and 10 (test) respectively. We choose the 10-class test set (600 images per class) from 10 different coarse labels (superclass). 100 images from each class construct the probe, the others work as the gallery. We use VGG16 [24] as our backbone architecture. The models are trained with batch size of 512, the learning rate starts from 0.1 and is divided by 0.4087 after each 10 epochs. The training is finished at 30 epochs. The mean Average Precision (mAP) is used to measure the performance.

2) RESULTS

First, we evaluate the effects of the setting of two hyper-parameters m and l of NMM loss. From Section Upper Bound of Inter-class Margin m, we get the upper bound of m is$\sqrt{\frac{1}{2}} \approx 0.71$. In addition, we find the performance of NMM is insensitive to intra-class margin l. Therefore, we set the l and m from 1.0 to 0.8, 0.2 to 0.7 respectively. If m is smaller than 0.2, the model fails to converge because it causes very large inter-class loss (see Section Upper Bound of Inter-class Margin m for detail). The results are shown in Table 1. With the increase of m, the mAP increases consistently, and achieves the best performance at $m = 0.7$. In addition, we can see that the performance is not sensitive to the setting of l. In this paper, l is set to 0.9.

We compare the proposed NMM with some state-of-the-art loss functions for open-set problem in CIFAR-100 dataset. The results are shown in Table 2. We use the same backbone network and all the loss functions select their best hyper-parameters ($\gamma = 10$ for COCO loss, $m = 4$ for

TABLE 1. The effects of the setting of margins $< l, m >$.

margin $< m, l >$	1.0	0.9	0.8
0.2	34.0%	32.0%	38.7%
0.3	33.6%	33.4%	34.1%
0.4	36.0%	39.7%	36.7%
0.5	40.7%	39.0%	40.2%
0.6	42.3%	42.9%	44.0%
0.7	46.6%	**47.3%**	46.4%

TABLE 2. Comparisons on CIFAR 100.

Method	mAP(%)
softmax loss	41.67
Center loss [11]	42.16
COCO loss [9]	44.08
SphereFace loss [3]	42.83
CosFace loss [12]	43.00
ArcFace loss [16]	44.2
NMM loss	**47.35**

TABLE 3. Comparisons on ImageNet.

Method	mAP(%)
softmax loss	28.7
Center loss [11]	31.3
COCO loss [9]	31.8
SphereFace loss [3]	32.9
CosFace loss [12]	33.3
ArcFace loss [16]	34.1
NMM loss	**34.8**

SphereFace, $\gamma = 5, \delta = 0.4$ for CosFace loss, $m = 0.5$ for ArcFace loss and learning rate 0.1 for all.) via grid searching.

We can see that our method significantly outperforms the other loss functions, which demonstrates the effectiveness of NMM for the open-set problem.

B. ImageNet (IMAGE CLASSIFICATION)

1) DATASET AND SETTINGS

ILSVR2012, a subset of ImageNet, contains 1000 classes and totally 1.3M training images. Similarly, to simulate the open-set problem, 900 classes with 1M images of training images are used for the model training and the remaining 100 classes are used to test the model. In the new test set, 100 images from each class are selected as the probe, the others are gallery. We use VGG16 [28] as our backbone architecture. In addition, all the images are reshaped to 256 × 256. The model is trained with batch size of 512, the learning rate starts from 0.1 and is divided by 0.4087 after each 10 epochs. The training is finished at 80 epochs.

2) RESULT

For a comparison, our NMM and 6 competitors (softmax + cross-entropy, center loss, COCO loss, SphereFace, Cosface, ArcFace loss) are trained using the same backbone architecture under their best hyperparameters. The classification results are shown in Table 3. From this table we can see that although more classes in ImageNet decrease the classification performance of the model compared with in CIFAR-100 dataset, NMM can learn more discriminative features of inter-class. Hence it outperforms other loss functions.

C. LFW (FACE RECOGNITION)

1) DATASETS

We use CASIA-WebFace [29] (removing the images of identities appearing in test sets) to train our models. CASIA-WebFace has 494,414 face images belonging to 10,575 different individuals. These face images are horizontally flipped for data augmentation. We use the well-known LFW benchmark [26] for test. LFW includes 13,233 face images from 5,749 different identities. The dataset contains faces with large variations in pose, expression and illuminations. We follow the ''unrestricted with labeled outside data'' protocol on the dataset. The score (metric) is computed by the cosine distance of two features.

2) PREPROCESSING AND SETTINGS

Face detection and landmark detection are performed by MTCNN [30]. The detected faces are aligned to 128 × 128.

For this experiment, we use ResNet50 [31] as our backbone architecture. The model is trained with batch size of 512 on four GPUs. The learning rate starts from 0.1 and is divided by 0.4087 after each 10 epochs. The training is finished at 80 epochs.

The face recognition results of ResNet50 with our NMM and 6 competitors are reported in Table 4. We can see that our NMM greatly outperforms the competitors.

We compare our NMM with state-of-the-art methods. Note that most existing face verification systems achieve very high performance with much bigger training data than ours. As shown in Table 4, the state-of-the-art methods use 200M images for FaceNet [32], 2.6M for DeeFR [33], and 4M for DeepFace [5], compared with 0.5M of our method. Despite this, our NMM achieves very competitive performance.

TABLE 4. Face recognition performance on LFW.

Method	Data	Accuracy(%)	mAP(%)
FaceNet [32]	200M	99.65	–
DeepFR [33]	2.6M	98.95	–
DeepFace [5]	4M	97.35	–
softmax loss	WebFace	97.45	97.48
Center loss [11]	WebFace	98.42	98.44
COCO loss [9]	WebFace	98.07	98.24
SphereFace loss [3]	WebFace	98.45	98.62
CosFace loss [12]	WebFace	98.84	99.01
ArcFace loss [16]	WebFace	99.01	99.23
NMM loss	WebFace	**99.04**	**99.24**

D. MSMT17 (PERSON RE-IDENTIFICATION)

1) DATASET AND SETTINGS

MSMT17 [27] contains 126,441 bounding boxes, 4,101 identities, which are significantly larger than the previous databases used for person re-identification. The ratio of training and test set is 1:3. The training set contains 32,621 bounding boxes of 1,041 identities, and the test set contains 93,820 bounding boxes of 3,060 identities. From the test set, 11,659 bounding boxes are randomly selected as query and the other 82,161 bounding boxes are used as gallery. We use ResNet50 [31] as our backbone architecture. The model is trained with batch size of 512 on four GPUs. The learning rate begins with 0.1 and is divided by 0.1 after each 30 epochs. The training is finished at 40 epochs.

2) RESULT

From Table 5, we can see our NMM outperforms another 6 competitors: softmax, Center loss, COCO loss, SphereFace loss, CosFace loss, ArcFace loss. We also compare NMM with other state-of-the-art methods reported on MSMT17 database: GoogLeNet [34], PDC [35] and GLAD [36]. Our NMM loss also works better than them, showing the effectiveness of our proposed method.

TABLE 5. Person re-id performance on MSMT17.

Method	mAP(%)
GoogLeNet [34]	23.0
PDC [35]	29.7
GLAD [36]	34.0
softmax loss	31.6
Center loss [11]	33.1
COCO loss [9]	33.4
SphereFace loss [3]	33.4
CosFace loss [12]	34.1
ArcFace loss [16]	34.6
NMM loss	**35.6**

V. CONCLUSION

In this paper, we propose a novel loss function NMM to guide the deep CNNs to learn highly discriminative features for boosting the performance of deep open-set image classification. Our NMM (1) *explicitly* maximizes the inter-class distance and minimizes the intra-class distance; (2) defines the margins of intra-class and inter-class; (3) does not need samples pairing while deep metric learning does. We provide the well-formed geometrical and theoretical interpretation to verify the effectiveness of the proposed NMM on generating strong feature representation, following by extensive experiments on various datasets. Furthermore, the upper bound of inter-class margin is innovatively determined by theoretical analysis. Our approach consistently outperforms many competitors: softmax, centre loss, COCO loss, SphereFace loss, CosFace loss and ArcFace loss across several benchmarks. Compared with the state-of-the-art methods, our method achieves very competitive performance.

In the future, we will explore the ways of automatically selecting the parameters for the margin and incorporating class-specific or sample-specific margins. In addition, we will extend the use of our method to other computer vision tasks.

REFERENCES

[1] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1563–1572.

[2] P. Oza and V. M. Patel, "C2AE: Class conditioned auto-encoder for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2307–2316.

[3] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6738–6746.

[4] Y. G. El-Yaniv, "Selective classification for deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4872–4887.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[6] D. O. Cardoso, J. Gama, and F. M. G. Franca, "Weightless neural networks for open set recognition," *Mach. Learn.*, vol. 106, nos. 9–10, pp. 1–21, 2017.

[7] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 507–516.

[8] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *Comput. Sci.*, 2017.

[9] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017.

[10] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "von Mises-Fisher mixture model-based deep learning: Application to face verification," *Comput. Sci.*, 2017.

[11] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[12] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018.

[13] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 539–546.

[14] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[15] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface L_2 hypersphere embedding for face verification," in *Proc. ACM Conf. Multimedia*, 2017, pp. 1041–1049.

[16] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019.

[17] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[18] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2317–2324, Nov. 2014.

[19] C. J. Parde, C. D. Castillo, M. Q. Hill, Y. I. Colon, S. Sankaranarayanan, J.-C. Chen, and A. J. O'Toole, "Deep convolutional neural network features and the original image," *CoRR*, vol. abs/1611.01751, 2016. [Online]. Available: http://arxiv.org/abs/1611.01751

[20] W. Wan, Y. Zhong, T. Li, and J. Chen, "Rethinking feature distribution for loss functions in image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9117–9126.

[21] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6398–6407.

[22] M. Karpusha, S. Yun, and I. Fehervari, "Calibrated neighborhood aware confidence measure for deep metric learning," 2020, *arXiv:2006.04935*. [Online]. Available: https://arxiv.org/abs/2006.04935

[23] I. Fehervari, A. Ravichandran, and S. Appalaraju, "Unbiased evaluation of deep metric learning algorithms," 2019, *arXiv:1911.12528*. [Online]. Available: https://arxiv.org/abs/1911.12528

[24] M. Kaya and H. S. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[26] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2007.

[27] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, 2014.

[29] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *Comput. Sci.*, 2014.

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[32] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 41.1–41.12.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014, *arXiv:1409.4842*. [Online]. Available: http://arxiv.org/abs/1409.4842

[35] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[36] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1709.04329

CHUNMEI MA received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2015. In 2015, she joined Nanyang Technological University (NTU), as a Visiting Scholar. She is currently a Lecturer with the School of Computer and Information Engineering, Tianjin Normal University, China. Her current research interests include intelligent computing, deep learning, and intelligent driving.

HUAZHI SUN received the Ph.D. degree in control theory and control engineering from the University of Science and Technology Beijing, Beijing, China, in 2008. He is currently a Professor and the Dean of the School of Computer and Information Engineering, Tianjin Normal University, China. His main research interests include artificial intelligence, pattern recognition, and machine learning.

JINQI ZHU received the Ph.D. degree in computer science from the University of Electronic Science and Technology of China (UESTC), China, in 2009. In 2013, she joined Nanyang Technological University (NTU), as a Visiting Scholar, under the supervision of Dr. Y. G. Wen. She is currently a Professor with the School of Computer and Information Engineering, Tianjin Normal University, China. Her main research interests include mobile computing and distributed computing.

LONG ZHANG received the Ph.D. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2014. He is currently a Professor with the School of Computer and Information Engineering, Tianjin Normal University, China. His current research interests include intelligent computing, deep learning, and signal processing.

BOJUE WANG received the B.E. degree in information security from Wuhan University, Hubei, China, in 2015. He is currently pursuing the M.S. degree in computer science and technology with Tianjin Normal University, Tianjin, China. His main research interests include reinforcement learning and model compression.

DONGHAO WU received the B.E. degree in computer science and technology from China Jiliang University, Zhejiang, China, in 2020. He is currently pursuing the M.S. degree in computer science and technology with Tianjin Normal University, Tianjin, China. His main research interest includes multimodal driving behavior recognition.

JINGWEI SUN received the B.S. degree in information and computing sciences from Jilin University, Jilin, China, in 2018. He is currently pursuing the M.S. degree in computer science with Tianjin Normal University, Tianjin, China. His main research interests include transfer learning and reinforcement learning.