J. Vis. Commun. Image R. 60 (2019) 51–58

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

SphereReID: Deep hypersphere manifold embedding for person re-identification ☆,☆☆

Xing Fan, Wei Jiang *, Hao Luo, Mengjuan Fei

Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 29 June 2018
Revised 21 November 2018
Accepted 4 January 2019
Available online 6 January 2019

Keywords:
Person re-identification
Classification
Feature embedding
CNN
Hypersphere

ABSTRACT

Many current successful Person Re-Identification (ReID) methods train a model with the softmax loss function to classify images of different persons and obtain the feature vectors at the same time. However, the underlying feature embedding space is ignored. In this paper, we use a modified softmax function, termed Sphere Softmax, to solve the classification problem and learn a hypersphere manifold embedding simultaneously. A balanced sampling strategy is also introduced. Finally, we propose a convolutional neural network called SphereReID adopting Sphere Softmax and training a single model end-to-end with a new warming-up learning rate schedule on four challenging datasets including Market-1501, DukeMTMC-reID, CHHK-03, and CUHK-SYSU. Experimental results demonstrate that this single model outperforms the state-of-the-art methods on all four datasets without fine-tuning or re-ranking. For example, it achieves 94.4% rank-1 accuracy on Market-1501 and 83.9% rank-1 accuracy on DukeMTMC-reID. The code and trained weights of our model will be released.

1. Introduction

Person re-identification is the task of identifying bounding box images of the same person from non-overlapping camera views. Given a probe image, we need to retrieve all images of the same person ID in gallery images.

Person re-identification has many practical applications such as video surveillance for public security, and thus attracts much research attention in the computer vision community. With the utilization of deep convolution neural networks (CNNs) [1] in recent years, ReID performance has made significant progress. However, some problems remain to be solved owing to the challenges of ReID, including changes in camera viewpoints, illumination changes, human pose variation and occlusion.

Most of the current ReID approaches can be categorized into two types: feature-based or metric-based. Extracting features from input images and seeking a metric for comparing these features across images are the two main components of person re-identification. Some handcrafted features such as scale-invariant feature transforms (SIFT) features [3,4] and local maximal occurrence(LOMO) features [5] have been used to represent a person's appearance.

With the success of deep learning, CNN-based methods have been proposed for ReID to automatically learn the feature representations from the training data. These methods [6–9] often model ReID as a classification problem and consider images from a specific person ID as a class. Then the softmax cross-entropy loss is applied to supervise the training procedure. Simultaneously, as a by-product, feature vectors before the last fully connected layer are extracted as the final image features. It corresponds with the intuition that when a feature vector can be used to classify a person ID correctly, it is a good representation of that person's appearance. However, without explicit constraints on the feature space distribution, the learned feature mapping may not be optimal. As shown in Fig. 1(a), there is no constraint on the distribution in the embedding space, which leads to a general spread.

To overcome the aforementioned drawbacks of feature-based works, metric-based methods [10–14] have been proposed to learn an embedding of the original images that satisfies some specified conditions. For example, triplet loss [2] requires the distance of samples from the same class to be less than that of samples from different classes by a pre-defined threshold, which pulls the instances of the same person closer and simultaneously pushes the instances belonging to different persons away from each other in the embedding space. Then the learned model is used for feature mapping of test images, and the extracted features can be com-

☆ This paper belongs to "5.1: feature extraction and representation", "5.13: deep neural networks" and "6.11: content based image retrieval" categories.
☆☆ This paper has been recommended for acceptance by Kuo-Liang Chung.
* Corresponding author.
E-mail address: jiangwei_zju@zju.edu.cn (W. Jiang).

https://doi.org/10.1016/j.jvcir.2019.01.010
1047-3203/

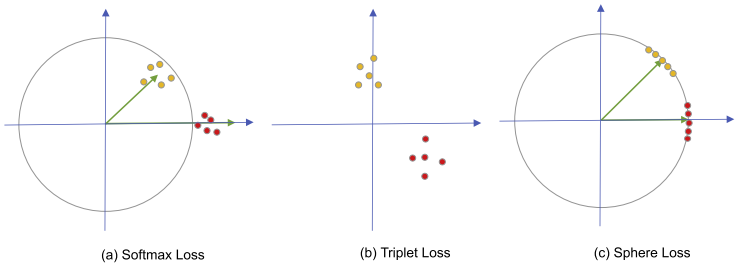

Fig. 1. Two-dimensional visualization of sample distribution in the embedding space supervised by (a) Softmax Loss, (b) Triplet Loss [2], and (c) Sphere Loss. Yellow and red points represent embedding features from two different classes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pared using the Euclidean distance criterion. However, the range of each dimension is from minus infinity to plus infinity, and the feature of each dimension only lies within a small interval, as shown in Fig. 1(b). Consequently, the target embedding space may not be fully utilized.

In this paper, we propose a novel metric-based person re-identification network called SphereReID, which adopts a new function called Sphere Loss to supervise the training process. Softmax cross-entropy is the basic loss function for the classification task. Despite the widespread use of softmax, whether it is the optimal loss function for classification is still uncertain. With the re-examination of softmax in the face recognition community [15–20], some valuable insights have been obtained. Motivated by their works, we adopt a modified softmax loss function called Sphere Loss, which classifies image samples from different persons and restrains the distribution of sample embeddings on a hypersphere manifold at the same time.

To the best of our knowledge, this is the first time a person image has been mapped onto a hypersphere manifold for person re-identification. To this end, feature normalization and weight normalization are introduced. After elimination of different norms, the classification will only rely on the angle between the embedding vector and the target class weight vector, which has a more clear geometric interpretation in the embedding space, as shown in Fig. 1(c), where embedding features lie on a hypersphere manifold. Compared with Euclidean space embedding, SphereReID maps images on the surface of a hypersphere, which limits the possible space distribution to a restricted angular space. Thus, the target embedding space can be fully exploited and we can train a network to classify images from different persons and simultaneously regulate the target embedding distribution. Furthermore, the implementation of Sphere Loss is simple and the code will be released.

One issue with person re-identification is that there are many datasets [21–29] and every labeled person has an indefinite quantity of images, thus sampling bias always exists. Further, some ReID datasets are image-based [21,24–26,28,29] whereas some are video-based [22,23,27] consisting of a lot of consecutive images frames, which makes the images per person ID even more diverse. A softmax supervised classification approach suffers from the sample amount bias and ends up with an inferior performance. In this paper, a balanced sampling strategy is introduced in the training process, and every mini-batch is generated by sampling a specific number of each person ID with a specific number of images which alleviates the sample amount bias problem.

With a new warming-up learning rate schedule, we train a single SphereReID model end-to-end without fine-tuning on four ReID datasets, and this single model outperforms the state-of-the-art methods on all the four datasets and achieves rank-1 accuracy 94.4% on Market-1501 [25], 83.9% on DukeMTMC-reID [28], 93.1% on CUHK03 [24] and 95.4% on CUHK-SYSU [26].

The contribution of our work is threefold:

– First, we introduce a new classification loss function called Sphere Loss modified from the softmax loss function, which can supervise the model to classify images of different persons and learn an embedding on a hypersphere manifold simultaneously.

– Second, a balanced sampling strategy is adopted to eliminate the sample amount bias and further facilitate the model performance without additional computational overhead. During training, a warming-up learning rate schedule is also used to bootstrap the network, which leads to a better convergence point.
– Finally, we propose a novel network called SphereReID adopting Sphere Loss. Extensive experiments on four datasets demonstrate the effectiveness of our proposed model.

2. Related works

Feature-based ReID. Some handcrafted features such as scale-invariant feature transforms (SIFT) features [3,4], Local Binary Patterns(LBP) features [30], and local maximal occurrence(LOMO) features [5] have been used to represent a person's appearance. With the rise of deep learning, automatically learning feature representations have been used for the ReID task and significant progress has been made as a result. Features extracted by a pre-trained CNN on a large annotated dataset, *e.g.*, ImageNet, have been proven to be strong off-the-shelf descriptors for various recognition tasks, and Matsukawa et al. [31] present CNN features for person re-identification fine-tuned on a pedestrian attribute dataset. To extract fine-grained part feature, Varior et al. [10] present a gate structure, while LSTM [32] is introduced in [11,33,34] to learn horizontal local features. Additonally, Sun et al. [35] use horizontal stripes and Li et al. [6] use a Spatial Transform Networks (STN) [36] subnet to localize the refined body parts and Zhao et al. [8] learn the parts automatically through a mask predictor. Furthermore, Yao et al. [37] represent different body parts by directly clustering feature maps based on the location of their maximum responses. As the human body is highly structured with known key points, external skeleton models have also been used for predicting different body regions in [9,38,39,7].

Metric-based ReID. Along with feature-based methods, there are some approaches to ReID that use metric learning, which formulate the person re-identification as a supervised distance metric learning problem. Traditional metric learning methods like the Keep It Simple and Straightforward Metric (KISSME) [40] and cross-view quadratic discriminant analysis (XQDA) [5] learn a transformation matrix of features. Nowadays, however, the community pays more attention to the loss function of a network. Instead of the softmax classification loss function, contrastive loss [41] is used to supervise a Siamese network in [11,10]. Motivated by FaceNet [42], a convolutional neural network used to learn an

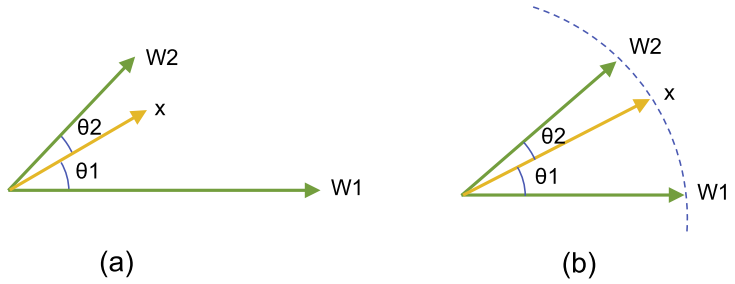

Fig. 2. Geometrical interpretation of (a) Softmax Loss and (b) Sphere Loss. Yellow arrows represent embedding feature vectors and green arrows represent class center weight vectors of two different classes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

embedding for faces, triplet loss [2] is used in [43,12] to optimize in the embedding space such that embedding features for the same identity are closer to each other than those of different identities. Cheng et al. [44] propose an improved triplet loss by introducing another pull term into the loss, penalizing large distances between positive embeddings. Quadruplet loss proposed in [14], adds another pull term for the distance between negative pairs, which can lead to a model with a larger inter-class variation and a smaller intra-class variation. Generation of samples of triplets or quadruplets will remain a challenge, as easy samples will lead to a degeneration and too difficult samples may result in gradient explosion. To solve this problem, Hermans et al. [13] propose a batch hard sampling strategy.

Other ReID methods. Xiao et al. [45] propose a domain guided dropout algorithm to improve the feature learning procedure for multiple ReID datasets. Geng et al. [46] introduce pairwise-consistent dropout for the pairwise verification loss layers, that is, each pair of compared training data points share the same dropout mask. AlignedReID [47] introduces a feature matching method to align different body parts. And DarkRank [48] shows that a powerful teacher model can significantly help the training of a smaller and faster student network for ReID. Re-ranking methods [49,50] can also be used to rearrange the original ranking list to further improve the accuracy. Generative adversarial nets (GAN) [51] have also been proven to be effective, and can also be exploited for the ReID task. PTGAN [52] proposes a Person Transfer Generative Adversarial Network (PTGAN) to bridge the domain gap between different datasets which relieves the expensive costs of annotating new training samples. Pose-normalization GAN (PN-GAN) [53] proposes a deep person image generation model for synthesizing realistic person images conditional on the pose.

Softmax re-examination in face recognition. The face recognition community has re-examined the meaning of softmax [15–19] and obtained valuable insights. Large-margin softmax (L-Softmax) loss is introduced in [15], and it maps the cosine value between feature vectors and the weight vector to a monotonically decreasing function with a large margin. The angular softmax (A-Softmax) loss proposed by [16] enables convolutional neural networks to learn angularly discriminative features and weight normalization is introduced. In [17–19], feature normalization is also applied, which makes the classification results only depend on the angle between the feature vector and weight vector.

3. Our approach

3.1. Softmax loss

In this section, we will discuss the meaning of the softmax loss function. Softmax is commonly used for the classification task. Given an input feature vector x_i with its corresponding label y_i, it can be formulated as follows:

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{z_{y_i}}}{\sum_{j=1}^{C}e^{z_j}} \tag{1}$$

where N is the number of training samples and C is the number of classes. z_j is activation of the j-th neuron in a fully connected layer with weight vector W_j and bias b_j. There are a total of C neurons, and each neuron outputs the score z_j of the corresponding j-th class. We fix the bias $b_j = 0$ for simplicity, and as a result, we can formulate z_j as follows:

$$z_j = W_j^T x = \|W_j\|\|x\|\cos\theta_j \tag{2}$$

where θ_j is the angle between W_j and x. As shown in Fig. 2(a), for an embedding feature vector x, and learnable weights W_1 and W_2 which serve as the class center, both the feature vector and weight vector influence the output scores. For a binary classification, when $z_1 > z_2$, the sample is classified into class 1, and class 2 otherwise. The decision boundary is as follows:

$$\|W_1\|\cos\theta_1 = \|W_2\|\cos\theta_2 \tag{3}$$

Eq. (3) shows that both the norm and angle influence the final decision. As shown in Fig. 3, there is an intersection area of class 1 and class 2, and thus samples of two classes cannot be distinguished only by the angle.

3.2. Sphere Loss

To eliminate the influence of norm and learn angularly discriminative features, we fix $\|W_j\| = 1$ and $\|x\| = 1$ by L2 normalization as follows:

$$W_j = \frac{W_j^*}{\|W_j^*\|}, x = \frac{x^*}{\|x^*\|} \tag{4}$$

where W_j^* and x^* are the original weight vector and feature vector.

In the original softmax loss function without normalization, when the angle between the feature vector and weight vector is the same, a sample tends to be classified into classes with a larger norm, which we call weight bias, and a sample with a larger norm tends to output a larger score, which we call feature bias. With the introduction of weight normalization and feature normalization, weight bias and feature bias are removed.

As shown in Fig. 2(b), after normalization, the weight vector and feature vector are all mapped onto a hypersphere manifold, and the classification results only depend on the angle between the feature vector and weight vector. For a binary classification, when $\cos\theta_1 > \cos\theta_2$, the sample is classified into class 1, and class 2 otherwise. The decision boundary is:

$$\cos\theta_1 = \cos\theta_2 \tag{5}$$

As shown in Fig. 3(b), compared with softmax, there is a clear decision boundary and classification results only depend on the angle.

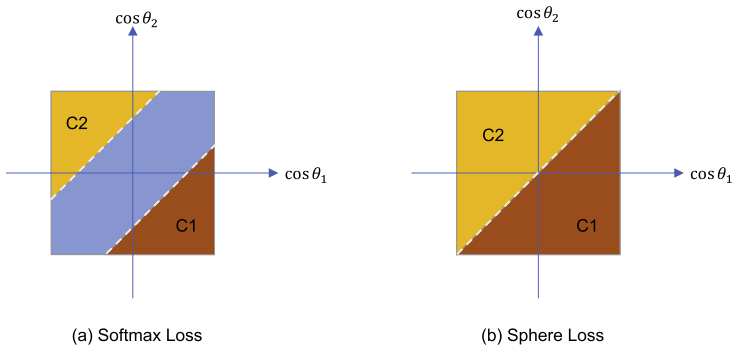

Fig. 3. The decision boundary of (a) Softmax Loss and (b) Sphere Loss. Samples lie within yellow area will be classified into class 2 and class 1 when samples lie within red area. Blue area means the class is uncertain because both angle and norm contribute to the decision. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Combining weight normalization and feature normalization, we also add a scale factor to control the temperature of the softmax function. Note that $\|W_j\| = 1$ and $\|x\| = 1$ and we then get the Sphere Loss:

$$L_{sphere} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{\sum_{j=1}^{C}e^{s\cos\theta_j}} \tag{6}$$

where s is the scale factor. In this paper, we use $s = 14$ for all experiments. Eq. (6) is similar to the normalized version of softmax loss (NSL) proposed in [18], but in [18], it is only an intermediate result proposed for the face recognition task and its effects are not fully exploited. Eq. (6) also matches the special case of additive margin softmax loss [17] and additive angular margin loss [19] when the margin is set to 0. A similar normalization also exists in NormFace [20].

With the supervision of Sphere Loss, we can learn an embedding on a hypersphere manifold, and different samples are discriminated by angles.

3.3. Balanced sampling strategy

A ReID datasets consist of images from different person where every person has an indefinite number of images. Usually, there is no constraint on the proportion of different persons in a mini-batch and training data is chosen randomly from all the training images. However, the training process suffers from an imbalance of data. The network trains more on a person with more images, while it trains less on a person with fewer images. Thus the model tends to fit more on the person with more images. However, there is an a priori that every person is of the same importance and should be treated equally.

Therefore, we introduce a balanced sampling strategy. To generate a mini-batch, we randomly choose P different persons without replacement, and for each person, we randomly choose K images. Thus there are a total of PK images in a mini-batch. For people with less than K images, we use sampling with replacement, and sampling without replacement otherwise. After all persons are sampled, we say that an epoch is considered done.

This balanced sampling strategy is similar to the strategy proposed in [13] for hard triplets mining, whereas in this paper, we use it to remove the imbalance of classes.

In this manner, for persons with more images we just ignore the over abundance of images, while for persons with fewer images we may use the same images multiple times. This approach guarantees that every person ID has the same number of instances.

3.4. SphereReID network

Combining Sphere Loss and the balanced sampling strategy, we propose a deep convolution neural network named SphereReID for the ReID task. As shown in Fig. 4, we use a ResNet-50 [54] network as the backbone network. After the last convolutional layer, a global average pooling follows to aggregate spatial information. Then we apply a batch normalization layer.

We also add a dropout layer as a regularizer, followed by a fully connected layer and another batch normalization layer. Now, we can apply weight normalization and feature normalization, and compute the Sphere Loss.

The global average pooling is a common operation in ReID area to aggregate features and reduce feature dimensions as in PCB + RPP

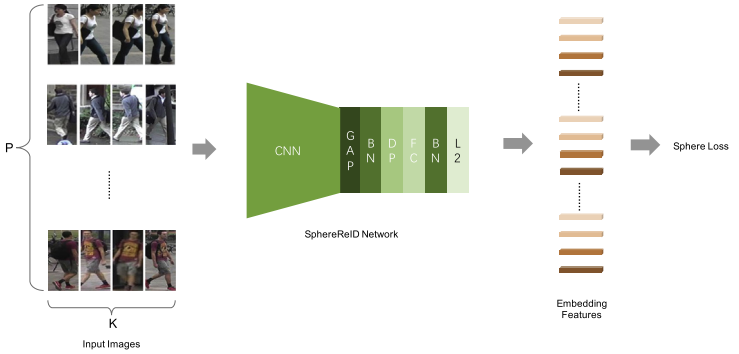

Fig. 4. The proposed SphereReID network structure. Inputs are a total of PK images generated by a balanced sampling strategy. After the last convolutional layer of the ResNet-50 [54] backbone, a global average pooling (GAP), batch normalization (BN), dropout (DP), fully connected layer (FC), batch normalization (BN), L2 normalization (L2) are follows respectively.

[35]. Batch normalization [55] can accelerate deep network training by reducing internal covariate shift and obtain a better performance as demonstrated in ArcFace [19]. Dropout is adopted to prevent overfitting as in MSCAN [6], then an FC is used to reduce features to the expected dimension. And L2 layer is part of our Sphere Loss in which features should be normalized to eliminate feature bias. Properly combining all those components together, we get a good design of network for ReID task, which is one of our contributions.

4. Experiments

4.1. Datasets

We conduct extensive experiments on four widely used ReID datasets: Market-1501 [25], DukeMTMC-reID [28], CUHK03 [24], and CUHK-SYSU [26].

Market-1501 contains 32,668 annotated bounding box images of 1501 labelled persons captured by five high-resolution cameras and one low-resolution camera. The dataset employes the Deformable Part Model (DPM) [56] as the pedestrian detector. A total of 751 persons are used for training.

DukeMTMC-reID is a subset of Duke-MTMC [57] for person re-identification. It contains 36,411 annotated bounding box images of 1812 different identities captured by eight high-resolution cameras. A total of 1404 identities are observed by at least two cameras, and the remaining 408 identities are distractors. The training set contains 16,522 images of 702 identities and the test set contains the other 702 identities.

CUHK03 contains 14,096 annotated bounding box images of 1467 identities. Each identity is observed by two disjoint camera views. There are two kinds of bounding boxes available: one is manually cropped and the other is automatically detected by DPM [56]. In this paper, we use the manually cropped version.

CUHK-SYSU containing 18,184 full images and 8432 identities. A total of 99,809 bounding box images are annotated from full images. The training set contains 11,206 full images and 5532 persons, whereas the test set contains 6978 full images of 2900 persons.

4.2. Implementation details

Our SphereReID model is built on the PyTorch framework. The backbone network is ResNet-50 [54] pre-trained on ImageNet and the original fully connected layer is discarded.

The inputs images are resized to 288×144 then randomly cropped to 256×128. The parameters P and K in the balanced sampling strategy are 16 and 4 respectively, as a result, a mini-batch size of 64 is used in our experiments.

We use the Adam optimizer with the default hyper-parameters ($\epsilon = 10^{-8}, \beta_1 = 0.9, \beta_2 = 0.99$). We set the initial learning rate to 10^{-3} and apply the decay schedule at epoch 80 and reduce the learning rate to 10^{-4}. At epoch 100, we reduce the learning rate again to 10^{-5}. The total number of training epochs for all conducted experiments is set to 140.

We also introduce a warming-up strategy to bootstrap the network, as shown in Fig. 5. We spend 20 epochs to linearly increase the learning rate from 5×10^{-5} to 10^{-3}. We think this strategy will help the network to initialize well before applying a large learning rate to optimize it. The experiment results are shown in the next section and demonstrate the effectiveness of this strategy.

4.3. Results of SphereReID

In this section, we go over different experiments settings and compare the rank-1 accuracies on Market-1501 [25], DukeMTMC-reID [28], CUHK03 [24], and CUHK-SYSU [26].

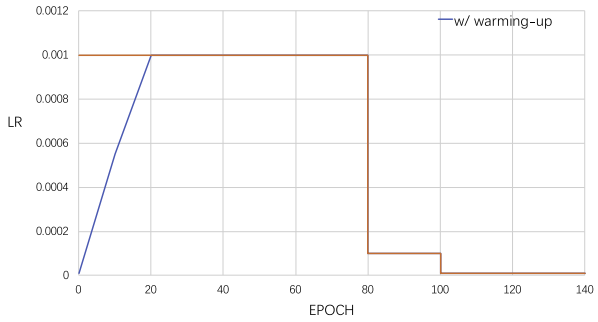

Fig. 5. Our learning rate schedules with or without warming-up. Blue one is with warming-up and the orange one is without warming-up. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Network structure and loss. After the last convolutional layer of the ResNet-50 [54] backbone, we have four different structures as follows: (A) global average pooling; (B) global average pooling, then a fully connected layer; (C) global average pooling, then a fully connected layer and a batch normalization; (D) global average pooling, batch normalization, dropout, fully connected layer and then a batch normalization again. The embedding feature size is 2048 for network-A and is 1024 for network-B, network-C and network-D. For network-D, the ratio of dropout is set to 0.5. Finally, L2 normalization is applied for all the networks.

The results are shown in Table 1. We can see that network-B is much better than network-A, which suggests that the additional fully connected layer can fuse input information and produce better embedding features. Network-C is also better than network-B, which demonstrates the effect of batch normalization. Network-D is the best and achieves 93.1% rank-1 accuracy on Market-1501 which combines the strength of batch normalization and dropout. Table 1 also shows that sphere Loss is clearly better than softmax.

The subsequent experiments all use the network-D structure.

Balanced sampling strategy. The balanced sampling strategy can guarantee that each training identity has the same number of instances and alleviates the imbalance of sample size. As shown in Table 2, when the model is trained with the balanced sampling strategy, the final performance is significantly boosted by a large margin, even with the exactly the same network structure. This strategy may be applied to a wider area of tasks, helping to eliminate the class bias in an imbalanced dataset.

Influence of warming-up. As shown in Table 2, with a warming-up process of the learning rate to help the network bootstrap, rather than applying a large learning rate from the beginning, the network can converge on a much better point. It is intuitive that when a network is initialized by weights pre-trained on ImageNet and never be used for the ReID task, a large learning rate may be inappropriate. The proposed fine-to-coarse then coarse-to-fine learning rate schedule can help set up a better initialization status and thus result in a better performance. The proposed warming-up strategy is not limited to the ReID task, and it may be applied to other areas to obtain a better optimizing result.

Ratio of dropout. We try three different ratios of dropout, and a network without dropout. Results are shown in Table 3. We can see that the networks with no dropout (when the ratio is 0) or with too much dropout are inferior to the network with a modest dropout ratio of 0.25. In the following experiments, we will fix the dropout ratio to 0.25.

Influence of the bias term. In the last fully connected layer, the bias term b can be set to 0 or learned automatically. We train two networks with and without the automatically learned bias term.

Table 1
Results of different network structures (with Balanced Sampling Strategy).

Network	Market-1501		DukeMTMC-reID		CUHK03		CUHK-SYSU	
	Sphere	Softmx	Sphere	Softmx	Sphere	Softmx	Sphere	Softmx
network-A	59.5	57.9	49.1	46.1	67.3	62.3	77.4	78.1
network-B	86.4	65.6	75.5	53.9	86.8	63.7	90.6	83.1
network-C	92.3	72.7	81.6	59.0	91.5	62.2	94.1	83.1
network-D	93.1	77.3	81.9	61.9	92.8	66.6	94.3	86.2

Table 2
The influence of balanced sampling strategy and warming-up strategy on Sphere Loss.

	Market-1501	DukeMTMC-reID	CUHK03	CUHK-SYSU
Balanced, w/ warming-up	93.1	81.9	92.8	94.3
Imbalanced, w/ warming-up	79.3	56.5	79.2	89.9
Balanced, w/o warming-up	77.1	64.4	80.6	88.9

Table 3
The influence of dropout ratio.

Ratio	Market-1501	DukeMTMC-reID	CUHK03	CUHK-SYSU
0	93.1	82.2	92.5	94.6
0.25	92.8	83.5	93.2	94.8
0.5	93.1	81.9	92.8	94.3
0.75	91.3	80.5	91.5	93.5

Table 4
The influence of bias term in the last fully connected layer.

	Market-1501	DukeMTMC-reID	CUHK03	CUHK-SYSU
w/ bias	93.7	83.9	92.6	94.9
w/o bias	92.8	83.5	93.2	94.8

Table 5
The influence of test image size. Images are resized to 288×144 with and without center crop of 256×128.

	Market-1501	DukeMTMC-reID	CUHK03	CUHK-SYSU
256×128	93.7	**83.9**	92.6	94.9
288×144	**94.4**	82.7	**93.1**	**95.4**

Table 6
Comparison with the State of the Art on Market-1501.

Method	rank-1	rank-5	rank-10	mAP
Spindle [7]	76.9	91.5	94.6	–
SVDNet [58]	82.3	92.3	95.2	62.1
PDC [39]	84.1	92.7	94.9	63.4
Mutual [59]	87.7	–	–	68.8
PSE [50]	87.7	94.5	96.8	69.0
PartLoss [37]	88.2	–	–	69.3
DPFL [60]	88.9	–	–	72.6
CamStyle [61]	89.5	–	–	71.6
GLAD [9]	89.9	–	–	73.9
HA-CNN [62]	91.2	–	–	75.7
Deep-Person [34]	92.3	–	–	79.6
PCB + RPP [35]	93.8	97.5	98.5	81.6
SphereReID	**94.4**	**98.0**	**98.7**	**83.6**

Table 7
Comparison with the State of the Art on CUHK-SYSU.

Method	rank-1	rank-5	rank-10	mAP
deep [26]	62.7	–	–	55.7
DLDP [63]	76.7	–	–	74.0
NPSM [64]	81.2	–	–	77.9
SphereReID	**95.4**	**98.6**	**98.9**	**93.9**

Table 8
Comparison with the State of the Art on DukeMTMC-reID.

Method	rank-1	rank-5	rank-10	mAP
SVDNet [58]	76.7	86.4	89.9	56.8
HA-CNN [62]	78.3	–	–	57.6
DPFL [60]	79.2	–	–	60.6
PSE [50]	79.8	89.7	92.2	62.0
HA-CNN [62]	80.5	–	–	63.8
Deep-Person [34]	80.9	–	–	64.8
PCB + RPP [35]	83.3	90.5	**92.5**	**69.2**
SphereReID	**83.9**	**90.6**	92.4	68.5

Table 9
Comparison with the State of the Art on CUHK03.

Method	rank-1	rank-5	rank-10
PartLoss [37]	82.8	96.6	98.6
GLAD [9]	85.0	97.9	99.1
DPFL [60]	86.7	–	–
Spindle [7]	88.5	97.8	98.6
PDC [39]	88.7	98.6	99.2
Deep-Person [34]	91.5	**99.0**	**99.5**
SphereReID	**93.1**	98.7	99.4

Results are shown in Table 4. We can see that the network with the bias term automatically learned performs slightly better than the network without the bias term.

From now on, we will refer this best network setting with a bias term as SphereReID network.

Test image size. In the training phase, we resize the image to 288×144, then randomly crop it to 256×128. In the testing phase, results of different image sizes are shown in Table 5. We can see that with larger input size, the performance is better on Market-1501, CUHK03, and CUHK-SYSU, and is worse on

Fig. 6. Some test examples. For every dataset, the first image in each row is the query image, and the rest 10 images are the most similar 10 gallery images found by SphereReID network ordered by similarity. As you can see, in Market-1501, all found images are very similar to query images. In the first row of DukeMTMC-reID, the occlusion is well-handled. In CUHK03 examples, there are pose changes, and SphereReID can still find the correct matches. In the first row in CUHK-SYSU, even there are not many visual clues, our method success to find the right person. Those examples demonstrate the efficiency of the proposed SphereReID.

DukeMTMC-reID. After examining images from the four datasets, we find that images from DukeMTMC-reID have a larger border background area. Thus, we use 256×128 test size for DukeMTMC-reID and 288×144 for the others.

4.4. Comparison with the state of the art

We compare SphereReID with the state of the art. As shown in Tables 6, 7, and 9, our single mode consistently outperforms the state of the art in terms of both accuracy and mAP, and it achieves 94.4% rank-1 accuracy on Market-1501. It is necessary to point out that no extra attributes, skeleton datasets, or models are used in our SphereReID network.

On DukeMTMC-reID, as shown in Table 8, PCB + RPP [35] obtains competitive results, but it is trained by a three stage process with fine-tuning. However, the proposed SphereReID is trained end-to-end without fine-tuning and is clearly better than PCB + RPP on Market-1501. Furthermore, SphereReID achieves all the results with a feature size of 1,024, while PCB + RPP uses a feature size of 12,288, which proves that our SphereReID features mapped onto a hypersphere manifold are more discriminative.

Some test examples are shown in Fig. 6, as you can see, our SphereReID can handle occlusions, illumination changes, and pose variations well, finding the most similar matches, which demonstrates the efficiency of the proposed SphereReID.

5. Conclusions

In this paper, we introduce a modified softmax loss function called Sphere Loss with weight normalization and feature normalization. We also propose a CNN network adopting Sphere Loss called SphereReID which can learn the feature embedding on a hypersphere manifold. We train the SphereReID end-to-end with the balanced sampling strategy and warming-up strategy and our single model outperforms the state of the art on all four datasets without re-ranking or fine-tuning.

To the best of our knowledge, this is the first network to learn a deep hypersphere manifold embedding for person re-identification, and the proposed SphereReID network demonstrates the effectiveness of this concept. We have provided a new idea for ReID and there are further improvements can be explored by the person re-identification community, for example, the addition of a margin term to increase inter-class variation and reduce intra-class variation.

And the proposed warming-up strategy can further boost the performance of deep neural network without extra computing overhead. It's very simple to implement and can be easily introduced into the training process. In this paper, we focus on SphereReID for person re-identification task, but it can also be used in other tasks, which remains for the computer vision community to explorer in the further.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgement

Our work was supported by the Public Projects of Zhejiang Province, China (No. LGF18F030002) and the National Natural Science Foundation of China (No. 61633019).

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[2] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (Feb) (2009) 207–244.
[3] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2528–2535.
[4] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: The IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3586–3593.
[5] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by Local Maximal Occurrence representation and metric learning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2197–2206.
[6] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[7] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[8] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: The IEEE International Conference on Computer Vision (ICCV), 2017.

[9] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, Glad: Global-local-alignment descriptor for pedestrian retrieval, in: ACM Multimedia, 2017, pp. 420–428.
[10] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision (ECCV), 2016, pp. 791–808.
[11] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: European Conference on Computer Vision (ECCV), Springer, 2016, pp. 135–153.
[12] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recogn. 48 (10) (2015) 2993–3003.
[13] A. Hermans, L. Beyer, B. Leibe, In Defense of the Triplet Loss for Person Re-Identification. ArXiv e-prints, 2017.
[14] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[15] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-Margin Softmax Loss for Convolutional Neural Networks. ArXiv e-prints (December 2016).
[16] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, SphereFace: deep hypersphere embedding for face recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[17] F. Wang, W. Liu, H. Liu, J. Cheng, Additive Margin Softmax for Face Verification. ArXiv e-prints (January 2018.
[18] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, W. Liu, CosFace: Large Margin Cosine Loss for Deep Face Recognition. ArXiv e-prints (January 2018).
[19] J. Deng, J. Guo, S. Zafeiriou, ArcFace: Additive Angular Margin Loss for Deep Face Recognition. ArXiv e-prints (January 2018).
[20] F. Wang, X. Xiang, J. Cheng, A.L. Yuille, NormFace: L2 Hypersphere Embedding for Face Verification, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017.
[21] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision (ECCV), 2008, pp. 262–275.
[22] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Image Analysis, 2011, pp. 91–102.
[23] T. Wang, S. Gong, X. Zhu, S. Wang, Person Re-Identification by Discriminative Selection in Video Ranking. ArXiv e-prints (January 2016).
[24] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: deep filter pairing neural network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: The IEEE International Conference on Computer Vision (ICCV), 2015.
[26] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, End-to-end deep learning for person search, 2016. arXiv preprint arXiv:1604.01850.
[27] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: a video benchmark for large-scale person re-identification, in: European Conference on Computer Vision (ECCV), 2016, pp. 868–884.
[28] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, 2017, arXiv preprint arXiv:1701.07717.
[29] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[30] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: European Conference on Computer Vision (ECCV), 2014, pp. 1–16.
[31] T. Matsukawa, E. Suzuki, Person re-identification using CNN features learned from combination of attributes, in: Proceedings of International Conference on Pattern Recognition (ICPR), 2016, pp. 2428–2433.
[32] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[33] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification, IEEE Trans. Image Process. 26 (July 2017) 3492–3506.
[34] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. ArXiv e-prints (November 2017).
[35] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond Part Models: Person Retrieval with Refined Part Pooling. ArXiv e-prints (November 2017).
[36] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.
[37] H. Yao, S. Zhang, Y. Zhang, J. Li, Q. Tian, Deep Representation Learning with Part Loss for Person Re-Identification. ArXiv e-prints (July 2017).
[38] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose Invariant Embedding for Deep Person Re-identification. ArXiv e-prints (January 2017).
[39] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven Deep Convolutional Model for Person Re-identification. ArXiv e-prints (September 2017).
[40] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2288–2295.
[41] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, The IEEE Conference on Computer Vision and Pattern Recognition, CVPR. Vol. 2, 2006, pp. 1735–1742.
[42] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.
[43] S. Khamis, C.H. Kuo, V.K. Singh, V.D. Shet, L.S. Davis, Joint learning for attribute-consistent person re-identification, in: European Conference on Computer Vision (ECCV), 2014, pp. 34–146.
[44] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[45] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[46] M. Geng, Y. Wang, T. Xiang, Y. Tian, Deep Transfer Learning for Person Re-identification. ArXiv e-prints (November 2016).
[47] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, J. Sun, AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. ArXiv e-prints (November 2017).
[48] Y. Chen, N. Wang, Z. Zhang, DarkRank: Accelerating Deep Metric Learning via Cross Sample Similarities Transfer. ArXiv e-prints (2017).
[49] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking Person Re-identification with k-reciprocal Encoding. ArXiv e-prints (2017).
[50] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A Pose-Sensitive Embedding for Person Re-Identification with Expanded Cross Neighborhood Re-Ranking. ArXiv e-prints (November 2017).
[51] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[52] L. Wei, S. Zhang, W. Gao, Q. Tian, Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. ArXiv e-prints (November 2017).
[53] X. Qian, Y. Fu, W. Wang, T. Xiang, Y. Wu, Y.G. Jiang, X. Xue, Pose-Normalized Image Generation for Person Re-identification. ArXiv e-prints (December 2017).
[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[55] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015.
[56] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[57] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision (ECCV), 2016, pp. 17–35.
[58] Y. Sun, L. Zheng, W. Deng, S. Wang, SVDNet for Pedestrian Retrieval, in: The IEEE International Conference on Computer Vision (ICCV), 2017.
[59] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[60] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2590–2600.
[61] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[62] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[63] A. Schumann, S. Gong, T. Schuchert, Deep Learning Prototype Domains for Person Re-Identification, 2016. arXiv preprint arXiv:1610.05047.
[64] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, S. Yan, Neural person search machines, in: The IEEE International Conference on Computer Vision (ICCV), 2017.