Article

Margin CosReid Network for Pedestrian Re-Identification

Xiao Yun ¹, Min Ge ¹, Yanjing Sun ^{1,*}, Kaiwen Dong ¹ and Xiaofeng Hou ²

¹ School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China; xyun@cumt.edu.cn (X.Y.); gemin@cumt.edu.cn (M.G.); dongkaiwen@cumt.edu.cn (K.D.)

² Wuxi Voicon Technology CO., LTD, 889 Zhenze Road, Wuxi 214000, China; houxiaofeng@voicon.cn

* Correspondence: yjcumt@cumt.edu.cn

Abstract: This paper proposes a margin CosReid network for effective pedestrian re-identification. Aiming to overcome the overfitting, gradient explosion, and loss function non-convergence problems caused by traditional CNNs, the proposed GBNeck model can realize a faster, stronger generalization, and more discriminative feature extraction task. Furthermore, to enhance the classification ability of the softmax loss function within classes, the margin cosine softmax loss (MCSL) is proposed through a boundary margin introduction to ensure intraclass compactness and interclass separability of the learning depth features and thus to build a stronger metric-based learning model for pedestrian re-identification. The effectiveness of the margin CosReid network was verified on the mainstream datasets Market-1501 and DukeMTMC-reID compared with other state-of-the-art pedestrian re-identification methods.

Keywords: pedestrian re-identification; CNN; softmax loss

Citation: Yun, X.; Ge, M.; Sun, Y.; Dong, K.; Hou, X. Margin CosReid Network for Pedestrian Re-Identification. *Appl. Sci.* **2021**, *11*, 1775. <https://doi.org/10.3390/app11041775>

Academic Editor: Davide Careglio

Received: 16 December 2020

Accepted: 12 February 2021

Published: 17 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Given a probe image (query), the goal of pedestrian re-identification is to search for images containing the same person in a gallery (database with training labels) under multiple nonoverlapping cameras [1]. Although it has played a supervisory role in public video-based pedestrian monitoring, the performance can be seriously affected by the presence or absence of obstructions and changes in target posture, camera angles, and illumination intensity.

To solve the above problems, an effective re-identification technology is expected to be able to identify person with the same identification under different environments, and distinguish between different people even if their appearances are similar. Therefore, a typical re-identification network should compose both feature learning ability for pedestrian's feature representation and distance metric learning ability for discriminative feature extraction [2,3]. Traditional studies on pedestrian re-identification are mostly based on handcrafted features, such as scale invariant feature transform (SIFT) features [4,5] and local maximum occurrence (LOMO) features [6], which are gradually replaced by neural networks. As the earliest simple neural network, the artificial neural network (ANN) is widely used in classification problems [7], but may decrease the algorithm accuracy and simultaneously increase the processing time [8]. The convolutional neural network (CNN), one of the typical feature extraction methods in deep learning, embeds the target feature space from the dataset automatically in the pedestrian re-identification problem. Compared to other CNN structures, ResNet50 is simpler and easier to train, which has been widely used in pedestrian re-identification methods [9]. Nevertheless, problems of overfitting, gradient explosion, and loss function non-convergence may occur in these networks, which are not very effective in dealing with complex re-identification tasks.

Besides a robust feature learning model in the re-identification network, an effective distance metric learning theory [10–13] is also the key to maximize the differences between

classes and minimize the differences within them in the re-identification problem. Some scholars proposed a support vector machine (SVM) that poses pedestrian re-identification as a metric-based ranking problem [14,15] by learning a ranking function parameterized by a weight vector, that is, sorting the positive sample pairs before the negative sample pairs. However, SVM can not provide fine-tuned features and may increase running time [8,16], which is gradually replaced by other metric-based re-identification methods such as contrastive loss [13] and triplet loss functions [10]. Traditional losses can only select simple and easily distinguishable sample pairs during training; therefore, Hermans et al. [11] proposed an effective metric-based method to solve hard sample selections but has convergence problems with loss function training and is time-consuming. Ahmed et al. [17] trained a pedestrian re-identification model on the softmax loss [18] function, which removes the classifier and uses the cosine similarity or Euclidean distance of the last layer of the network for distance query. The softmax loss function can solve the convergence problem without considering the minimum sample batch size and succeeds in performing classification between classes but has difficulties distinguishing the differences within the classification.

Considering the simple and effective performance of CNN and the success of the softmax loss function in the task of pedestrian re-identification, we combine CNN (i.e., feature-based reidentification) and softmax (i.e., metric-based reidentification) to achieve better results. However, the problems involving CNN and softmax which we mentioned above may hinder the successful performance of this combination. Therefore, we propose the margin CosReid network in this paper to extract more discriminative features from CNN and to introduce a margin parameter in softmax to solve their respective problems for effective pedestrian reidentification. The main contributions of this paper are summarized as follows:

- The proposed feature extraction model GBNeck is added behind the backbone network ResNet50 to avoid the overfitting and slow convergence problems caused by traditional CNNs and thus achieve a stronger generalization and more discriminative feature learning models.
- The proposed margin cosine softmax loss (MCSL) introduces a boundary margin parameter that can maximize differences between classes and minimize those within classes simultaneously to deal with outside interferences strongly.
- Our method was tested on Market-1501 [1] and DukeMTMC-reID [19] and performs superior compared to state-of-the-art pedestrian re-identification methods.

2. Related Work

Existing researches aiming to address the pedestrian re-identification problem mainly focus on different aspects of the issue such as developing robust feature learning models and designing discriminative metrics. In this section, we review several related works briefly.

2.1. Feature-Based Learning Re-Identification Methods

Traditional feature-based learning methods that handle the appearance variations in re-identification are mostly based on handcrafted features, such as SIFT [4,5] and LOMO [6]. However, handcrafted features are difficult to achieve satisfactory performance with the growing size of the re-identification dataset. As deep learning develops, automatically learning feature representations from the training data have been applied in the re-identification task and the network structure becomes much more complex [20]. Zheng et al. [21] applied CNNs to use the pedestrians' identification as training labels in the re-identification network. However, relying only on the identification information when training the network model often creates an overfitting problem, which leads to a model generalization inability. Thus, Lin et al. [22] proposed attribute-person recognition to combine the loss of pedestrians' identification with those of the multiple attribute identifications as training supervision. Sun et al. [23] proposed a visibility-aware part

model (VPM) to capture fine-grained learning features. Moreover, to take full advantage of the strengths of different features, Yang et al. [24] formulated a method for exploring more diverse discriminative visual features using a class activation maps (CAM) augmentation multibranch model and a novel penalty mechanism. Additionally, Chen et al. [25] introduced a pair of complementary attention modules and regularized network diversity, which was able to learn more discriminative features and reduce correlations. To prevent pedestrian pose variations from affecting the re-identification accuracy, Zheng et al. [19] first applied the generated adversarial network (GAN) to pedestrian re-identification tasks and proved its effectiveness, and Qian et al. [26] proposed a post-normalized GAN based on it. To narrow the gap between different datasets, Wei et al. [27] proposed a human transfer GAN to reduce the expensive data annotation on the datasets. Sun et al. [28] conducted a special study on the change in pedestrians' viewpoints to obtain diverse feature information. Nevertheless, performance may fail by using above methods when the whole body features are not representative in complex scenes. Therefore, Liu et al. [29] proposed a spatial and temporal features mixture model to make full use of the information from individual human body parts.

2.2. Metric-Based Learning Re-Identification Methods

Studies that focus only on feature-based learning may have difficulties on distinguishing similar appearance, which can be solved through metric learning methods. Traditional metric-based re-identification methods, such as cross-view quadratic discriminant analysis (XQDA) [6] and keep it simple and straightforward metric (KISSME) [30], learn feature subspace with resolution abilities. Kalarani et al. [8] combined SVM and ANN for data mining analysis to cope with the low accuracy and high processing time of ANN. To increase the interclass distance, Tang et al. [16] used hinge loss together with SVM to enhance the generalization ability of the model. To force the distance between dissimilar sample pairs to be greater than that between similar sample pairs, Cheng et al. [31] proposed an improved triplet loss, and Chen et al. [12] proposed quaternion loss and introduced new constraints based on it. As the network selects samples randomly, samples that are too simple will lead to poor generalization ability of the model, and samples that are too complex will lead to gradient explosion of the network during training. Therefore, selecting triplet and quaternion samples is still a major problem [11,32]. Hermans et al. [11] proposed hard sample mining to solve the poor generalization ability caused by random sample selection. Ahmed et al. [17] applied the softmax loss function to pedestrian re-identification by using the cosine similarity or Euclidean distance of the last layer of the network for distance query. Due to its strong classification ability, the softmax function remains widely used in applications of multiclassification, such as image classification, face recognition, and object detection [33,34]. Guo et al. [35] trained both softmax and center losses in ANN to classify the extracted spectral features. This combination makes the intraclass features more compact and the interclass features more expandable, but results in parameter redundancy simultaneously. Thus, Wen et al. [36] combined center loss with softmax loss to supervise CNN to minimize the deep feature distance between classes. To approximate the classification boundary in the metric-based learning re-identification, a linearization method [37] was proposed on a deep network and to enhance the robustness of the loss function by the activation function. The margin of the loss function in this method is imposed on any network layer but may lead to the expansion of parameters and longer training time.

3. Margin CosReid Network

The overall framework of the proposed margin CosReid network is shown in Figure 1. First, the dataset is inputted to the basic skeleton ResNet50. After that, the proposed feature extraction model GBNeck is added for feature extraction. Finally, the pedestrian descriptor is obtained.

Figure 1. The framework of the proposed margin CosReid network. The first dotted box is the basic skeleton of ResNet50 for dataset feature extraction, in which the blue bars represent the residual structure. Then, the global average pooling (GAP) layer is introduced. The second dotted box is the proposed GBNeck module. Finally, 512-dimensional feature vectors are obtained and inputted into the loss function for the final result.

3.1. Proposed GBNeck

In this paper, we propose the fine-tuned GBNeck model to extract deep pedestrian features. First, to increase the size of the feature map and obtain higher resolution features, the proposed GBNeck removes the last downsampling layer of the backbone network ResNet50. Then, to reduce the number of parameters and integrate the global spatial information, a global average pooling (GAP) layer is added to replace the fully connected (FC) layer behind ResNet50, and the kernel is pooled with a size of 16×8 into 1×1 to obtain a 2048-dimensional feature vector. Second, each neuron in the total connective layer is fully connected with all the neurons in the former layer to integrate the classified local information in the pooling layer.

Furthermore, the batch normalization (BN) layer can speed up training and minimize exploding gradients [38,39]. Thus, we introduce the BN layer in our GBNeck and experimentally find that it can also improve the generalization ability of the model. Then, the dropout layer is introduced to avoid the overfitting problem, improve the generalization performance, and play a regularization role in the training process [40]. Finally, another BN and FC layer is added again as the discriminative descriptor to focus the network on the input image and reduce the image distortion caused by external factors to obtain 512-dimensional feature vectors for person re-identification. The proposed GBNeck model can achieve faster convergence, stronger generalization, and more discriminative feature learning abilities in the training process.

3.2. Margin Cosine Softmax Loss (MCSL)

The traditional softmax loss function is widely used as a supervision mechanism in pedestrian re-identification because of its brilliant performance in discriminating features between classes but has difficulties distinguishing the differences within classes [41]. To solve this problem, this paper proposes the margin cosine softmax loss (MCSL) to normalize the weight and feature vector and introduces a boundary margin parameter m for maximizing the difference between classes while minimizing those within classes to embed the pedestrian features deeper.

3.2.1. Softmax Loss

This section first introduces the common classification loss function, namely, the softmax loss function. Given an input feature vector x_i of the i -th training sample and the corresponding label y_i , the traditional softmax loss function is expressed as

$$L_{softmax} = -\frac{1}{N} \sum_{i=1}^N \log p_i = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{f_{y_i}}}{\sum_{j=1}^C e^{f_j}}, \quad (1)$$

where p_i denotes the posterior probability conditioning that x_i is correctly classified. N and C denote the size of the training sample and the number of categories, respectively. f denotes the activation of the fully connected layer, including the weight vector $W \in \mathbb{R}^{d \times n}$ and the offset $B \in \mathbb{R}^n$. In Equation (1), $f_j = W_j^T x_i + B_j$ and $f_{y_i} = W_{y_i}^T x_i + B_{y_i}$, where W_j and W_{y_i} denote the j -th and y_i -th columns of the weight vector W , respectively. In this paper, B is set as 0 [42], and then the activation f_j is computed as

$$f_j = W_j^T x = \|W_j\| \|x\| \cos \theta_{j,i}, \quad (2)$$

where $\theta_{j,i}$ ($0 \leq \theta_{j,i} \leq \pi$) is the angle between the weight vector W_j and the feature vector x .

3.2.2. Cosine Softmax Loss

L-softmax [43] does not consider the imbalance of sample distribution; thus, we normalize $\|W_j\| = 1$ [33,42,44] so that each category is treated relatively equally during training. The feature vector $\|x_i\|$ is also L2 normalized and scaled to s , making the posterior probability p_i rely only on the cosine value for an improvement in the resolution ability. Then, the improved loss function, named the cosine softmax loss (CSL), is defined as [42]

$$L_{csl} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s \cos(\theta_{y_i,i})}}{\sum_{j=1}^C e^{s \cos(\theta_{j,i})}}. \quad (3)$$

In Equation (3), the features learned in the CSL space can be separated and thus classified correctly.

3.2.3. Proposed Margin Cosine Softmax Loss

Although features between different classes can be well distinguished, those within the same class cannot be separated using CSL. Based on the work in [43], SphereFace [42] normalizes the weight vector, and the multiplicative margin compresses the sample features to a smaller space and at the same time reduces the monotone interval of the cosine function, which leads to the difficulty of optimization. To solve this problem, both ArcFace [33] and CosFace [44] introduce additive margins and only have margins on the items corresponding to the real class labels. To maximize the distance between classes, we also consider the need to introduce additive margins between different categories at the denominator and propose the margin cosine softmax loss (MCSL) defined as

$$L_{mcsl} = -\frac{1}{N} \sum_i \log \frac{e^{s(\cos(\theta_{y_i,i})-m)}}{\sum_{j=1}^C e^{s(\cos(\theta_{j,i})-m)}}, \quad (4)$$

subject to

$$W = \frac{W^*}{\|W^*\|}, x = \frac{x^*}{\|x^*\|}, \cos(\theta_{j,i}) = W_j^T x_i, \quad (5)$$

where N and C denote the numbers of training sample batches and the dataset category, respectively. x_i denotes the normalized feature vector of the i -th sample corresponding to label y_i , and W_j denotes the weight vector of the category j . $\theta_{y_i,i}$ and $\theta_{j,i}$ denote the angle between x_i and the weight vectors W_{y_i} or W_j , respectively. Compared to the exp-normalize trick [45], MCSL maps the pedestrian descriptor to the cosine space through a cosine function that has intrinsic consistency with softmax. The hyperparameter m is the boundary margin by introducing which MCSL can make classification stricter through controlling the cosine boundary size from $\cos(\varphi_1) > \cos(\varphi_2)$ to $\cos(\varphi_1) - m > \cos(\varphi_2)$. φ_i ($i = 1, 2$) denotes the angle between the feature vectors x_i , and C_i ($i = 1, 2$) denotes the class label of x_i . x_i is determined to belong to class C_1 by $\cos(\varphi_1) - m > \cos(\varphi_2)$ in MCSL for significant performance in distinguishing features within class. Moreover, the binary classified MCSL can also be expanded to other multiple classification problems.

3.3. Loss Comparison

In this section, the decision boundaries among classes C_1 and C_2 of the three loss functions softmax, CSL, and MCSL are shown in Figure 2.

Figure 2. Comparison of decision boundaries of (a) softmax, (b) cosine softmax loss (CSL), and (c) margin cosine softmax loss (MCSL) functions under binary classification. Green and orange regions represent the decision areas of classes C_1 and C_2 , respectively. The yellow region represents the overlapping decision area. Purple and blue dotted lines represent the decision boundaries of classes C_1 and C_2 , respectively.

First, in the traditional softmax loss function, overlapping decision area (denoted by the yellow region in Figure 2a) exists between the decision boundaries of classes C_1 and C_2 (denoted by purple and blue dotted lines, respectively) defined as $\|W_1\| \cos(\varphi_1) = \|W_2\| \cos(\varphi_2)$, so that decisions in this overlapping area cannot be made on which class it belongs to.

Based on softmax, CSL performs L2 normalization on the weight vectors W_1 and W_2 , thus making the decision boundary a constant ($\cos(\varphi_1) = \cos(\varphi_2)$). As shown in Figure 2b, boundaries of classes C_1 and C_2 are connected together, so that the decisions on or closed to the connected line are also unable to be made on their associated classes. CSL performs well in simple pedestrian classification problems because there is no large overlapping area compared to traditional softmax, but is still vulnerable to outside interference with similar pairs of negative samples and confuses two pedestrians with different identities.

To solve the above problems, the boundary threshold m is introduced in cosine space in the proposed MCSL algorithm. The boundary conditions of the MCSL are defined as

$$C_1 : \cos(\varphi_1) \geq \cos(\varphi_2) + m, \quad (6)$$

$$C_2 : \cos(\varphi_2) \geq \cos(\varphi_1) + m, \quad (7)$$

In Equations (6) and (7), the feature vector belongs to class C_1 only if the minimum value of $\cos(\varphi_1)$ is greater than or equal to the maximum value of $\cos(\varphi_2)$, and vice versa. Because the decision boundaries of classes C_1 and C_2 are separated far away from each other as shown in Figure 2c, decisions are easier to be made in the MCSL function. Therefore, the interclass difference becomes larger, and the intraclass difference becomes more compact, which is strong enough to deal with outside interferences to learn more powerful distinguishing features.

To compare more depth about the three loss functions, we conduct an experiment on their class activation maps in Figure 3. Color temperature means importance in producing gradients associated with the target identities, indicating that warmer colors and sharper outlines demonstrate better performances. The rough outlines of results are drawn with pink dotted curves. In Figure 3a, the outline of MCSL has sharper and more similar shape with that of the person in the input image than those of softmax and CSL. In Figure 3b,c, the color in the outline of MCSL is warmer than those of softmax and CSL. Moreover, red boxes in Figure 3 denote the backgrounds that are mistakenly identified with target in

outlines of both softmax and CSL, whereas not in that of MCSL, which indicates that MCSL outperform others in class activation mapping.

Figure 3. An example of class activation maps of the three input person images (a–c) on softmax, CSL, and MCSL functions. Warmer color with higher value. The rough outlines and the backgrounds that are mistaken to the target are marked with pink dotted curves and red boxes, respectively.

4. Experiment

4.1. Datasets

In our experiments, all the algorithms are tested on the two datasets: Market-1501 [1] and DukeMTMC-reID [19]. The Market-1501 dataset consists of 32,668 images of 1501 labeled persons from six cameras, one of which has a lower resolution. There are 751 identities in the training set and 750 identities in the testing set. The DukeMTMC-reID dataset contains 36,411 images of 1812 persons of eight high-resolution camera views. A total of 16,522 images of 702 persons are randomly selected from the dataset as the training set, and the other 702 persons are divided into the testing set, which includes 2228 query images and 17,661 gallery images.

We demonstrate the effectiveness of the proposed method by using the cumulative matching characteristics (CMC) at rank-1, rank-5, and rank-10 and the mean average precision (mAP) on the standard dataset. Considering reidentification as a ranking problem, rank- n ($n = 1, 5, 10$) denotes the probability that the first n identities ranking in candidate re-identification lists have the correct result, in which rank-1 is the most important. The mAP calculates the mean value of average precisions (APs) for all queries, which considers both precision and recall of an algorithm, thus providing a more comprehensive evaluation [1]. The higher the values of rank-1, rank-5, rank-10, and mAP, the better of the re-identification performances.

4.2. Experimental Results and Analysis

Our approach is implemented using the PyTorch framework with GTX TITAN X GPU, Intel i7 CPU, and 128 GB memory. The skeleton network ResNet50 is pretrained in ImageNet [46]. The sizes of the input images are adjusted to 256×128 , the batch size is set to 32, and the training iteration number is set to 60. The initial learning rate is set to 0.1 and reduces to 0.01 after 20 iterations. The weight attenuation is set to 0.0005, and the momentum term is set to 0.9.

4.2.1. GBNeck

Table 1 demonstrates the efficiency of GBNeck in Figure 1. Based on the baseline ResNet50, Net-A adds only global average pooling (GAP) layer; Net-B adds a fully connected (FC) layer based on Net-A; Net-C adds a batch normalization (BN) layer based on Net-B; and the proposed GBNeck adds a dropout layer, an FC layer, and a BN based on Net-C. The embedded feature size is 2048 for Net-A, 1024 for Net-B and Net-C, and 512 for GBNeck, where the dropout is set to 0.5. As shown in Table 1, the precision of Net-B is higher than Net-A because the FC layer includes the key image information to obtain more discriminative features. Net-C performs even better than Net-B, which shows the

effective generalization ability of the BN layer. GBNeck achieves the best rank-1 to 91.0% in Market-1501 by combining dropout and BN layers.

Table 1. Rank-1 and mAP of the four network frameworks. Bold fonts denote the highest values.

Network	Market-1501						DukeMTMC-reID					
	Softmax		CSL		MCSL		Softmax		CSL		MCSL	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
Net-A	62.9	34.6	74.3	48.3	77.3	51.6	56.8	31.6	61.1	35.6	65.1	37.8
Net-B	75.1	51.0	85.9	66.7	86.8	68.4	69.7	43.5	76.0	51.6	77.9	59.8
Net-C	78.4	54.4	87.3	67.3	88.9	70.5	76.7	59.1	78.3	57.5	78.6	60.0
GBNeck	81.7	57.9	87.4	69.2	91.0	75.9	77.8	59.9	79.3	61.0	81.1	64.6

In the proposed GBNeck, the last downsampling layer is removed from the backbone network ResNet50 for higher spatial resolution to bring significant improvement [47]. Two sets of experiments were performed on the influence of the downsampling layer, and the results are shown in Table 2. In this section, we demonstrate the effects of this removal.

Table 2. Influences on rank-1 and mAP of the downsampling layer. w/and w/o denote that the downsampling layer is and is not removed, respectively. Bold fonts denote highest values.

	Market-1501		DukeMTMC-reID	
	rank-1	mAP	rank-1	mAP
w/ downsampling layer	91.0	75.9	81.1	64.6
w/o downsampling layer	86.3	68.3	76.4	56.7

In addition, we also conduct an experiment to show the mean training loss curves from 5 runs of MCSL with the backbone ResNet50 and GBNeck on the Market-1501 dataset. We can see from Figure 4 that the GBNeck has faster convergence compared to ResNet50.

Figure 4. The mean training loss curves from 5 runs of MCSL with the backbone ResNet50 and GBNeck on the Market-1501 dataset. The horizontal and vertical coordinates represent the iteration number and the loss value, respectively.

4.2.2. The Proposed MCSL

In this section, we conduct experiments to discuss the efficiency of the proposed margin cosine loss (MCSL) function. Figure 5 demonstrates the mean training loss curves from 5 runs of the three functions, i.e., softmax, CSL, and MCSL optimized by Adam [48] and stochastic gradient descent (SGD) on the Market-1501 dataset. The training curves

of softmax loss function optimized by Adam and SGD decline slowly after the 20th and 15th iterations, respectively. Comparatively, the convergence of CSL in the early stage is accelerated due to the cosine space mapping, but it converges slowly in after the 25th iterations. The curve of MCSL optimized by Adam vibrates greatly in the final training stage, showing a clear confrontation between MCSL and Adam [49]. MCSL optimized by SGD can suppress the inconsistency and achieve a high accuracy from Figure 5, so we use SGD to optimize our training model in this paper.

Figure 5. The mean training loss curves from 5 runs of softmax, CSL, and MCSL optimized by Adam and SGD on the Market-1501 dataset. The horizontal and vertical coordinates represent the iteration number and the loss value, respectively.

In this paper, experiments are performed on the selection of the loss parameters s and m in Equation (4). The scaling parameter s in the proposed MCSL controls the dependence degree of the loss to the cosine space, which cannot be adaptively learned to prevent slow network convergence and relatively difficult optimization. As shown in Figure 6a, we found that if s is set too small, the loss will decrease slowly and may not even converge. Therefore, s is set to a larger value in this paper for better performance and training loss reduction. In addition, the margin parameter m can also be selected, as shown in Figure 6b. We set m to 0.35 in our paper because of the highest accuracy in both the Market-1501 and DukeMTMC-reID datasets.

Figure 6. Influence of parameters. Subfigure (a) indicates loss convergence with different scaling factors s on Market-1501. The horizontal and vertical coordinates represent the iteration number and the loss value, respectively. Subfigure (b) indicates the accuracy of the MCSL with different margin parameters m on Market-1501 and DukeMTMC-reID. The horizontal and vertical coordinates represent the iteration number and the accuracy, respectively.

4.2.3. Visualization Result

The visualization result of the proposed method is shown in Figure 7. People in different rows represent different pedestrian sample images in the Market-1501 dataset. Images in the first column denote the query graph, and the retrieval images are sorted according to the cosine similarity from 1 to 10. As seen from the sorting order, most of the retrieval images are selected correctly, but there are still some incorrect images marked with a red number. This failure probably occurred because of the insufficient image information collected from single view cameras.

Figure 7. Pedestrian retrieval results on the Market-1501 dataset. The first column in the black box shows five different identity query images. The other columns show the results sorted according to the similarity in which the sort orders are marked by serial numbers. The green numbers represent the retrieved images that are the same identities as the query, and the red numbers represent those without the same identities.

4.3. Comparison with State-of-the-Art Methods

The proposed re-identification method has been compared with other existing state-of-the-art methods, and the comparison results are shown in Tables 3 and 4. On the Market-1501 dataset, the single-person search rank-1 and mAP of our method reach 91.0% and 75.9%, respectively. On DukeMTMC-reID, the single-person search rank-1 and mAP of our method reach 81.1% and 64.61%, respectively, which are much higher than those of other methods. Compared with PSE [50], the rank-1 and mAP of our method without any personal posture information increase 2.3% and 6.9%, respectively, on Market-1501 and 1.3% and 2.61%, respectively, on DukeMTMC-reID. Furthermore, compared with PNGAN [26] using GAN to generate pedestrian images, rank-1 and mAP of our method increase 0.6% and 3.3%, respectively, on Market-1501 and 7.5% and 11.4%, respectively, on DukeMTMC-reID.

To show the improvements of the re-ranking method and the multi-query mode used on rank-1, rank-5, rank-10, and mAP, we conduct comparison experiments on the performances of the proposed MCSL with single-query mode (MCSL+single-query), MCSL with reranking (MCSL+RK), MCSL with multi-query mode (MCSL+multi-query), and MCSL with both reranking and multi-query mode (MCSL+RK+multi-query). The experimental results are shown in Tables 3 and 4, indicating that using re-ranking and multi-query mode can improve the performances of our method.

Table 3. Comparisons of rank-1, rank-5, rank-10, and mAP with different methods on the Market-1501 dataset. Bold fonts denote the highest values.

Method	Rank-1	Rank-5	Rank-10	mAP
PIE [51]	78.7	90.3	93.6	53.9
SVDNet [52]	82.3	92.3	95.2	62.1
LSRO [19]	83.4	-	-	66.1
PDC [53]	84.1	92.7	94.9	63.4
APR [22]	84.3	93.2	95.2	64.7
DML [54]	87.7	-	-	68.8
PSE [50]	87.7	94.5	96.8	69.0
Camstyle [55]	88.1	-	-	68.7
PNGAN [26]	89.4	-	-	72.6
AWTL [56]	89.5	-	-	75.7
Deep KPM [57]	90.1	-	-	75.3
MaskReID [58]	90.4	-	-	75.4
MCSL+single-query	91.0	96.6	98.0	75.9
MCSL+RK	92.6	95.8	97.1	87.9
MCSL+multi-query	93.8	97.9	98.8	83.0
MCSL+RK+multi-query	94.2	98.1	98.9	88.5

Table 4. Comparisons of rank-1, rank-5, rank-10, and mAP with different methods on the DukeMTMC-reID dataset. Bold fonts denote the highest values.

Method	Rank-1	Rank-5	Rank-10	mAP
LSRO [19]	67.7	-	-	47.1
APR [22]	70.7	-	-	51.9
PAN [59]	71.6	83.9	90.6	51.5
PNGAN [26]	73.6	-	-	53.2
SVDNet [52]	76.7	86.4	89.9	56.8
MaskReID [58]	78.8	-	-	61.9
PSE [50]	79.8	89.7	92.2	62.0
AWTL [56]	79.8	-	-	63.4
Deep KPM [57]	80.3	-	-	63.2
HA-CNN [60]	80.5	-	-	63.8
MCSL	81.1	90.6	93.3	64.6
MCSL+RK	82.9	91.5	94.2	73.4

5. Conclusions

In this paper, we proposed a margin CosReid network for end-to-end pedestrian re-identification with respect to feature and metric-based learning combinations. On the one hand, the feature extraction model GBNeck was proposed to overcome the overfitting and slow convergence problems of traditional CNNs, which is a typical model of feature-based learning. On the other hand, the margin cosine softmax loss (MCSL) was proposed by introducing a boundary margin parameter that can maximize the difference between classes and minimize those within classes simultaneously. Our margin CosReid network can not only achieve a stronger generalization and more discriminative feature learning models, but can also deal with outside interferences strongly for the metric-based learning model. It was tested on Market-1501 and DukeMTMC-reID and had superior performance compared to state-of-the-art pedestrian re-identification methods in accuracy and robustness.

Although a common phenomenon in deep network, the accuracy saturation problem of the proposed MCSL still affects the accuracy enhancement. In the future work, we plan to introduce a deep residual learning to decrease accuracy saturation and obtain an even higher accuracy.

Author Contributions: X.Y. designed the study, analysed data, and wrote the paper and the revisions. M.G. performed the experiments and analysed data. Y.S. contributed to refining the ideas, and is responsible for the research team. K.D. performed the experiments and analysed data. X.H. collected and analysed the data. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by [the Natural Science Foundation of Jiangsu Province] grant number [BK20180640], [the National Natural Science Foundation of China] grant number [61902404, 512918914, 62001475, 51804304, 61771417, 51734009, 61873246], and [the State Key Research Development Program] grant number [2016YFC0801403].

Data Availability Statement: Not applicable.

Acknowledgments: This work is supported by the Natural Science Foundation of Jiangsu Province (BK20180640), the National Natural Science Foundation of China (61902404, 512918914, 62001475, 51804304, 61771417, 51734009, and 61873246), and the State Key Research Development Program (2016YFC0801403).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
2. Jin, X.; Lan, C.; Zeng, W.; Chen, Z.; Zhang, L. Style normalization and restitution for generalizable person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3140–3149.
3. Song, W.; Zheng, J.; Wu, Y.; Chen, C.; Liu, F. Discriminative feature extraction for video person re-identification via multi-task network. *Appl. Intell.* **2021**, *51*, 788–803. [\[CrossRef\]](#)
4. Zhao, R.; Ouyang, W.L.; Wang, X.G. Person re-identification by salience matching. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2528–2535.
5. Zhao, R.; Ouyang, W.L.; Wang, X.G. Unsupervised salience learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3586–3593.
6. Liao, S.; Hu, Y.; Zhu, X.Y.; Li, S.Z. Person re-identification by Local Maximal Occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
7. Chen, L.S.; Liu, C.H.; Chiu, H.J. A neural network based approach for sentiment classification in the blogosphere. *J. Inf.* **2011**, *5*, 313–322. [\[CrossRef\]](#)
8. Kalarani, P.; Brunda, S.S. An efficient approach for ensemble of SVM and ANN for sentiment classification. In Proceedings of the IEEE International Conference on Advances in Computer Applications, Coimbatore, India, 24 October 2016; pp. 99–103.
9. Wang, H.; Sun, S.; Zhou, L.; Guo, L.; Li, C. Local feature-aware siamese matching model for vehicle re-identification. *Appl. Sci.* **2020**, *10*, 2474. [\[CrossRef\]](#)
10. Fan, X.; Jiang, W.; Luo, H.; Mao, W.; Yu, H. Instance hard triplet loss for in-video person re-identification. *Appl. Sci.* **2020**, *10*, 2198. [\[CrossRef\]](#)
11. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1–10.
12. Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond triplet loss: A deep quadruplet network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1320–1329.
13. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 1735–1742.
14. Jin, P.X.; Yang, B. Active learning with re-sampling for support vector machine in person re-identification. In Proceedings of the International Conference on Machine Learning and Cybernetics, Tianjin, China, 14–17 July 2013; pp. 597–602.
15. Zhang, Y.; Li, B.; Lu, H. Sample-specific SVM learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1278–1287.
16. Tang, Y.C. Deep learning using linear support vector machines. *arXiv* **2013**, arXiv:1306.0239.
17. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
18. Bishop, C.M. Pattern recognition and machine learning. *J. Electron. Imaging* **2006**, *16*, 140–155.
19. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by GAN improve the person re-identification baseline in Vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3774–3782.
20. Wei, L.; Wei, Z.; Jin, Z.; Yu, Z.; Huang, J.; Cai, D.; He, X.; Hua, X. SIF: Self-inspired feature learning for person re-identification. *IEEE Trans. Image Process.* **2020**, *29*, 4942–4951. [\[CrossRef\]](#) [\[PubMed\]](#)

21. Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q. Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3346–3355.
22. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.U.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* **2019**, *95*, 3346–3355. [\[CrossRef\]](#)
23. Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; Sun, J. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 393–402.
24. Yang, W.; Huang, H.; Zhang, Z.; Chen, X.; Huang, K.; Zhang, S. Towards rich feature discovery with class activation maps augmentation for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1389–1398.
25. Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; Wang, Z. ABD-net: Attentive but diverse person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8350–8360.
26. Qian, X.; Fu, Y.; Xiang, T.; Wang, W.; Qiu, J.; Wu, Y.; Jiang, Y.; Xue, X. Pose-normalized image generation for person re-identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 650–667.
27. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer GAN to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
28. Sun, X.; Zheng, L. Dissecting person re-identification from the viewpoint of viewpoint. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 608–617.
29. Liu, J.; Sun, C.; Xu, X.; Xu, B.M.; Yu, S.Y. A spatial and temporal features mixture model with body parts for video-based person re-identification. *Appl. Intell.* **2019**, *49*, 3436–3446. [\[CrossRef\]](#)
30. Kostinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
31. Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; Zheng, N. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1335–1344.
32. Fan, X.; Jiang, W.; Luo, H.; Fei, M. SphereReID: Deep hypersphere manifold embedding for person re-identification. *J. Vis. Commun. Image Represent.* **2019**, *60*, 51–58. [\[CrossRef\]](#)
33. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. ArcFace: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4685–4694.
34. Wang, X.; Shrivastava, A.; Gupta, A. A-fast-RCNN: Hard positive generation via adversary for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3039–3048.
35. Guo, A.J.X.; Zhu, F. Spectral-spatial feature extraction and classification by ANN supervised with center loss in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 99. [\[CrossRef\]](#)
36. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 499–515.
37. Elsayed, G.F.; Krishnan, D.; Mobahi, H. Large margin deep networks for classification. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018; pp. 850–860.
38. Zhang, C.; Nguyen, T.; Sah, S.; Ptucha, R.; Loui, A.; Salvaggio, C. Batch-normalized recurrent highway networks. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 640–644.
39. Laurent, C.; Pereyra, G.; Brakel, P.; Zhang, Y.; Bengio, Y. Batch Normalized Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 2657–2661.
40. Hinton, G.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. *Comput. Sci.* **2012**, *3*, 212–223.
41. Jie, S.; He, X.; Qing, L.; Yu, Y.; Xu, S.; Peng, Y. A new discriminative feature learning for person re-identification using additive angular margin softmax loss. In Proceedings of the UK/ China Emerging Technologies, Glasgow, UK, 21–22 August 2019; pp. 1–4.
42. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. SphereFace: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6738–6746.
43. Liu, W.G.; Wen, Y.; Yu, Z.; Yang, M. Large-margin softmax loss for convolutional neural networks. In Proceedings of the International Conference on International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; p. 7.
44. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. CosFace: Large margin Cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
45. Vieira, T. Exp-Normalize Trick. 2014. Available online: <https://timvieira.github.io/blog/post/2014/02/11/exp-normalize-trick/> (accessed on 16 February 2020).
46. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–16 December 2012; pp. 1097–1105.

47. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 480–496.
48. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.
49. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **2020**, *22*, 2597–2609. [[CrossRef](#)]
50. Sarfraz, M.S.; Schumann, A.; Eberle, A.; Stiefelhagen, R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 420–429.
51. Zheng, L.; Huang, Y.; Lu, H.; Yang, Y. Pose-invariant embedding for deep person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 4500–4509. [[CrossRef](#)] [[PubMed](#)]
52. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. SVDNet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3820–3828.
53. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3980–3989.
54. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H.C. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
55. Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camera style adaptation for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5157–5166.
56. Ristani, E.; Tomasi, C. Features for multi-target multi-camera tracking and re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6036–6046.
57. Shen, Y.; Xiao, T.; Li, H.; Yi, S.; Wang, X. End-to-end deep kronecker-product matching for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6886–6895.
58. Qi, L.; Huo, J.; Wang, L.; Shi, Y.; Gao, Y. A mask based deep ranking neural network for person retrieval. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo, Shanghai, China, 8–12 July 2019; pp. 496–501.
59. Zheng, Z.; Zheng, L.; Yang, Y. Pedestrian alignment network for large-scale person re-identification. *Proc. IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 3037–3045. [[CrossRef](#)]
60. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294.