symmetry

MDPI

Article

Learning Large Margin Multiple Granularity Features with an Improved Siamese Network for Person Re-Identification

Da-Xiang Li [1,2], Guo-Yuan Fei [1,*] and Shyh-Wei Teng [3]

1 School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; www_ldx@163.com
2 Ministry of Public Security Key Laboratory of Electronic Information Application Technology for Scene Investigation, Xi'an 710121, China
3 Faculty of Science & Technology, Federation University Australia, Gippsland, VIC 3842, Australia; imgcsiu@163.com
* Correspondence: fgy519597702@gmail.com

Received: 1 December 2019; Accepted: 25 December 2019; Published: 3 January 2020

Abstract: Person re-identification (Re-ID) is a non-overlapping multi-camera retrieval task to match different images of the same person, and it has become a hot research topic in many fields, such as surveillance security, criminal investigation, and video analysis. As one kind of important architecture for person re-identification, Siamese networks usually adopt standard softmax loss function, and they can only obtain the global features of person images, ignoring the local features and the large margin for classification. In this paper, we design a novel symmetric Siamese network model named Siamese Multiple Granularity Network (SMGN), which can jointly learn the large margin multiple granularity features and similarity metrics for person re-identification. Firstly, two branches for global and local feature extraction are designed in the backbone of the proposed SMGN model, and the extracted features are concatenated together as multiple granularity features of person images. Then, to enhance their discriminating ability, the multiple channel weighted fusion (MCWF) loss function is constructed for the SMGN model, which includes the verification loss and identification loss of the training image pair. Extensive comparative experiments on four benchmark datasets (CUHK01, CUHK03, Market-1501 and DukeMTMC-reID) show the effectiveness of our proposed method and its performance outperforms many state-of-the-art methods.

Keywords: person re-identification; multiple granularity features; Siamese Multiple Granularity Network; multi-channel weighted fusion loss

1. Introduction

Person re-identification is a crucial task in video analytics scenarios and it received more and more attention on computer vision field [1,2]. Person re-identification, as a core technology in video analysis, aims to determine whether the objects appearing in the non-overlapping view belong to the same person. Although the researchers have made great efforts to deal with this problem, it still has challenges because of large variations in viewpoints, backgrounds, illuminations and poses. As we can see in Figure 1, there are some hard samples from baseline datasets and those difficulties usually appear in realistic camera networks.

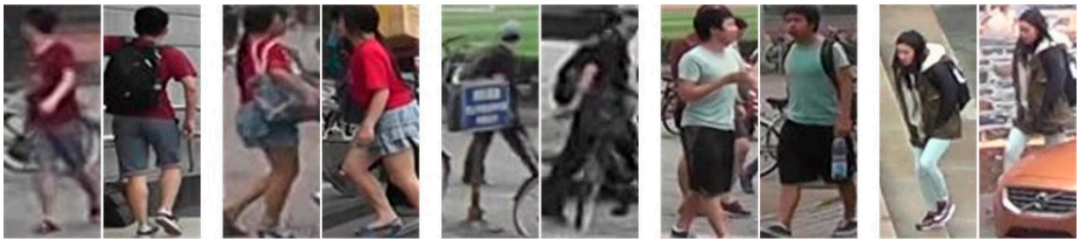
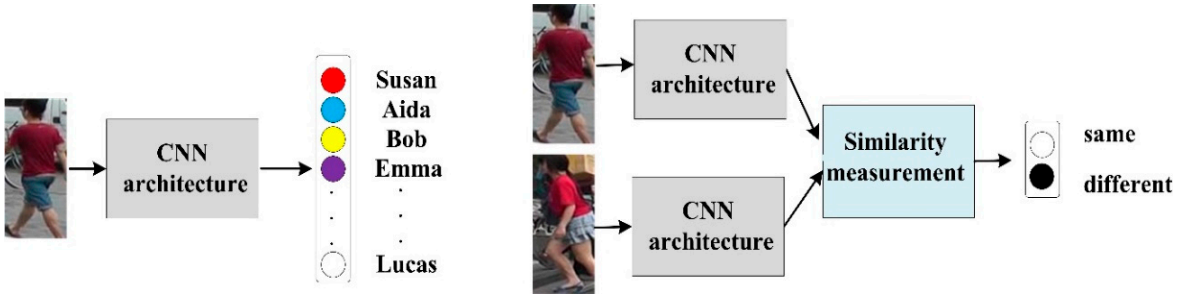


Figure 1. Example pairs of images from baseline person re-identification datasets. Every two adjacent images represent the same person. Analysis of these images suffered from much larger differences indicates person re-identification is challenging.

In order to realize person re-identification, the traditional research work mainly includes two aspects, namely feature extraction [3–6] and metric learning [7,8]. In feature extraction module, different pedestrian image descriptors are adopted to obtain discriminative information of pedestrian images. In metric learning module, there are various kind of distance metrics that are designed to find a suitable embedding space, in which the distance between similar data is pushed as close as possible while the distance between different data is pulled as far as possible.

Considering the success of deep learning in image classification problems, many researchers have applied it to person re-identification [9,10]. According to the differences in model structure, related algorithms can be divided into two categories as shown in Figure 2, namely the CNN-based identification model and Siamese based verification model. In the CNN-based identification model, the images in the training set and their labels are fed into CNN during the training processing. In order to obtain the discriminative features of pedestrian images, various loss functions are designed to take full advantage of the label information of the images, such as cross entropy loss [11], OIM (online instance matching) loss [12] etc. However, in the identification model, the problem is that it usually only uses the global information and ignores the local information of the images. In addition, the similarity metric between image pairs is not considered during model training [9–14]. Therefore, a Siamese-based verification model is proposed, which can judge whether the pedestrians in the two input images are the same person [15,16]. Compared with the identification model, the verification model constructs a loss function between the pairs of training images, and its focus is only on the similarity metric between the image pairs (that is, maximizing the similarity between positive pairs while minimizing the similarity between negative pairs as much as possible). In this case, this kind of model does not make use of the label information of the images during the training phase, which accounts for the final features of images not having the character of margin maximization for classification.



(**a**) Identification model (**b**) Verification model

Figure 2. The difference between the CNN-based identification model and the Siamese-based verification model. Identification models take one image as input and predict its identity while verification models take a pair of images as input and determine whether they belong to the same person or not.

In order to overcome the problems of the two models mentioned above during person re-identification, we fuse the two models together and design a new Siamese network model named

Siamese multiple granularity network (SMGN) in this paper. The backbone CNN of the SMGN is composed by two feature extraction branches, i.e., global and local feature extraction branches. In the proposed SMGN model, four identification loss functions and a verification loss function are designed to obtain the final multi-channel weighted fusion (MCWF) loss function. Therefore, SMGN is able to combine the advantages of identification model and verification model, and the final extracted multiple granularity features of pedestrian images have the characteristic of margin maximization for classification, namely large margin multiple granularities (LMMG) features. As a result, the algorithm based on SMGN can improve the performance of person re-identification.

The contributions of our work are threefold as follows:

- We propose a novel symmetric Siamese network model called SMGN, the backbone CNN of which is composed by two branches, i.e., a local branch and a global branch. Compared with the traditional Siamese network model, SMGN can obtain LMMG features of person images, including local features and global features, which would be of great benefit to person re-identification.
- By fusing the verification and the identification information, a new MCWF loss function is designed for the SMGN model. Compared with traditional cross entropy loss, MCWF loss function takes into account decision boundary information in identification channels, so LMMG features extracted from SMGN can be guaranteed to have the character of margin maximization for classification.
- We implement extensive experiments on four challenging person re-identification datasets (i.e., CUHK01 [17], CUHK03 [9], Market-1501 [18] and DukeMTMC-reID [19]). The experimental results show the proposed method achieves better results than the state-of-the-art methods.

The remainder of our paper is organized as follows: some related works are reviewed in Section 2. The structure of our proposed model and implementation details are presented in Section 3. Extensive comparative experiment results on four benchmark datasets are shown in Section 4, followed by conclusions drawn in Section 5.

2. Related Work

In this section, some previous works related to person re-identification are described simply.

2.1. Hand-Crafted Feature-Based Person Re-ID

The majority of traditional methods related to person re-identification pay close attention to two basic modules, i.e., feature extraction and metric learning. For feature extraction, several effective appearance cues attempt to build a robust feature representation. For example, Farenzena et al. [3] proposed symmetry-driven accumulation of local features (SDALF) to characterize pedestrian images, which are robust to image scale and illumination variations. SDALF consist of three kind of features, i.e., weighted color histograms, maximally stable color regions (MSCR) and recurrent high-structured patches (RHSP). In order to obtain discriminative features of pedestrian images, Local Maximal Occurrence representation (LOMO) is proposed by Liao et al. [4], which includes Scale Invariant Local Ternary Pattern (SILTP) descriptor and two scales of the local HSV histogram. Similarly, Yang et al. [5] utilized salient a Salient Color Name-Based Color Descriptor (SCNCD) that takes advantage of the robustness of color names to illumination to characterize pedestrian images. To further improve the performance, a Hierarchal Gaussian descriptor (GOG) was discussed in [6] that models the region as a set of multiple Gaussian distributions in which each Gaussian represents the appearance of a local patch.

For metric learning, different distance metrics have been proposed to learn a suitable metric space, in which the distance between the same pedestrian are kept as close as possible while the distance between different pedestrians are kept as far as possible. Representative metric methods include XQDA [4], KISSME [7], MLAPG [8] etc. Liao et al. [4] utilized cross-view quadratic discriminant analysis to learn a low dimensional subspace in which all the features have a character of discrimination;

meanwhile, a QDA metric is introduced. In [7], the decision on whether an image pair is similar or not is expressed as a likelihood ratio test. The pairwise difference method is adopted, and the difference space is a zero-mean Gaussian distribution. A logistic metric learning approach with the positive semi-definite (PSD) constraint and an asymmetric sample weighting strategy is derived in [8].

2.2. Deep Learned Feature-Based Person Re-ID

Previous hand-crafted descriptors and metric learning methods have made limited performance on person re-identification. Hence, many researchers tended to utilize CNN-based methods to solve person re-identification problems. Some work [20–22] shows that CNN have a great potential on image classification, object recognition, natural language processing etc. For person re-identification, Li [9] proposed a filter pairing neural network based on CNN that learn filter pairs to encode photometric transforms. Ahmed [10] proposed an enhanced deep learning framework to compute cross-input neighborhood differences and patch summary features. With the popularity of Siamese network, many works have devoted to using it to improve performance. Zheng [11] proposed a unit network that combines identification model and verification model, which learns a discriminative embedding and a similarity measurement simultaneously. Wu [13] proposed a Siamese attention structure based on joint learning spatiotemporal video representation and its similarity measurement. Chung [14] presented a two-stream convolutional neural network, in which each stream is a Siamese network. This architecture can learn spatial and temporal information separately. Benefiting from powerful deep networks, they achieved many state-of-the-art results on person re-identification.

2.3. Loss Function-Based Person Re-ID

As a supervised signal, loss functions play an important role in CNN models. For person re-identification, there are various loss functions have been proposed, such as cross entropy loss [15,23,24], binary classification loss [25,26], contrastive loss [27], center loss [28], triplet loss [29] etc. Cross entropy loss is the most popular used loss function for person re-identification, and it consider identification labels as supervised signals for reducing classification error; binary classification loss considers the deep network as a two-class model, classifying positive and negative sample from the image pair. As for contrastive loss, the Euclidean distance between two features is calculated directly by it, in order to minimize the distance between positive samples and punish the distance between negative samples when it is less than the threshold; center loss forces the similar image features into closing to their corresponding class center to reduce the intra-class variance, but it ignores pushing the distance among inter-class; Triplet loss makes the distance between positive pairs smaller than negative pairs, in other words, the distance between positive samples is pushed as close as possible while the distance between negative samples is pulled as far as possible. In addition, some loss functions based on softmax loss achieve state-of-art performance in face recognition. Liu et al. [30] proposed L-Softmax by adding angular constraints to each identity to improve the discrimination of pedestrian image features. A-Softmax [31] improves L-Softmax by normalizing the weights to learn angularly discriminative features. In addition, feature normalization is applied in [32], so that the classification results only depend on the angle between the feature vector and weight vector.

3. The Proposed Method

In this section, we first present the structure of the proposed SMGN model. Then we describe the MCWF loss function for the SMGN model. Thirdly, the training mechanism and cosine distance used in the testing phases are introduced. Finally a brief algorithm flow is concluded.

3.1. The Structure of SMGN

The overall network architecture of the proposed SMGN model is illustrated in Figure 3. It is essentially a five-channel Siamese model (including four identification channels and a verification channel), which takes a pair of person images as input. In the proposed SMGN model,

ResNet-50 is adopted as its backbone CNN because it has a competitive performance in person re-identification [10–12,16]. In order to use the local and global features to represent pedestrian images simultaneously, the subsequent part after res_conv4_1 block is divided into two independent branches in ResNet-50, namely, global and local feature extraction branches. Table 1 lists the settings of both the local and global branches. "Map Size" denotes the size of output feature maps from each branch. "Dimension" denotes the dimensionality of features for the output representations. "Feature" denotes the symbols for the output feature.

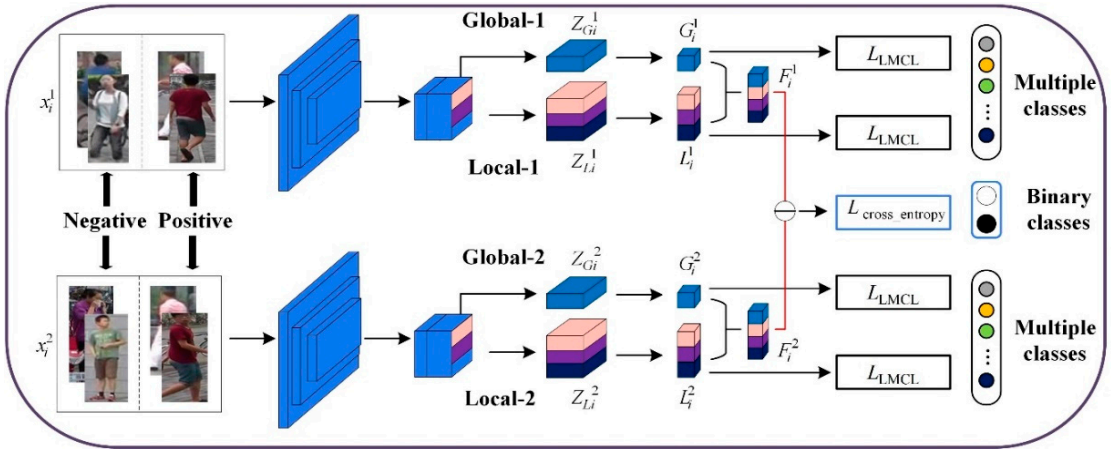


Figure 3. The framework of the proposed Siamese Multiple Granularity Network (SMGN).

Table 1. Comparison of the settings for four branches in SMGN.

Branch	Map Size	Dimension	Feature
Global-1	12×4	256	G_i^1
Global-2	12×4	256	G_i^2
Local-1	24×8	256×3	L_i^1
Local-2	24×8	256×3	L_i^2

As shown in Figure 3, "Global-1" and "Global-2" are global extraction branches while "Local-1" and "Local-2" are local extraction branches. In the global branch, down-sampling with a stride-2 convolution layer is adopted in res_conv5_1 block to address the problem that the output feature maps are sensitive to the location in the input images. After that, we perform global max-pooling (GAP) [33] operation on the corresponding output feature map. Meanwhile, batch normalization [34] and ReLU are introduced to accelerate the training and perform feature reduction respectively. In each global branch, we reduce 2048-dim features $Z_{G_i^j} \middle| j = 1, 2$ to 256-dim features $G_i^j \middle| j = 1, 2$. Different from the global branch, no down-sampling operations are adopted in the res_conv5_1 block. In this way, the appropriate areas of reception fields can be reserved for the local feature in the local feature extraction branch. Furthermore, we divide the feature maps into three uniform parts horizontally and the same following operations are conducted as the global feature extraction branch to obtain the local features of pedestrian images.

3.2. Multiple Granularity Features

During the training phase, we assume that an image pair $(x_i^1, x_i^2, l_i^1, l_i^2)$ is fed into SMGN, where x_i^1 and x_i^2 represent the first and second image in i-th image pair, and l_i^1 and l_i^2 denote the corresponding label of x_i^1 and x_i^2. The proposed SMGN can produce their descriptors from global branches and local branches. For the first image x_i^1, we can obtain its global features G_i^1 from the branch "Global-1" and its local features L_i^1 from the branch "Local-1". Similarly, we can get global features G_i^2 and local features

L_i^2 of the second image x_i^2. Finally, we concatenate global features and local features together to obtain the final representation of x_i^1 and x_i^2 through Equation (1) as follows:

$$\begin{cases} F_i^1 = [L_i^1, G_i^1] \\ F_i^2 = [L_i^2, G_i^2] \end{cases} \tag{1}$$

where F_i^1 and F_i^2 represent the multiple granularity features of the person image x_i^1 and x_i^2 respectively, which include both global information and local information from the corresponding images.

3.3. Multi-Channel Weighted Fusion Loss

To further improve the discriminability of multiple granularity features for person re-identification, we design a multi-channel weighted fusion (MCWF) loss function which include identification loss and verification loss in four identification channels and a verification channel.

3.3.1. Identification Loss

In the proposed SMGN model, there are four identification channels. For each identification channel, we introduce a new classification loss called large margin cosine loss (LMCL) [35] to make multiple granularity features have the character of margin maximization for classification.

In the traditional softmax loss function, different classes can be distinguished by maximizing the posterior probability of the ground-truth class. We assume that the i-th feature vector and its label are v_i and l_i respectively, then we can write the traditional softmax loss function as follows:

$$Loss_{\text{softmax}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{y_{l_i}}}{\sum_{j=1}^{C} e^{y_j}} \tag{2}$$

where N and C represent the number of training samples and classes respectively. Here, y_j represents the activation value of the j-th neuron in a fully connected layer with a weight vector W_j and a bias b_j. Relatively, there are C neurons in total, and the output of neurons represents the score that v_i belongs to the corresponding class. For the purpose of simplicity, we fix the bias $b_j = 0$, and then y_j can be computed by:

$$y_j = W_j^T v = \|W_j\|\|v\| \cos\theta_j \tag{3}$$

where v represents an input feature vector and θ_j is the angle between W_j and v.

In order to perform feature learning effectively, we fix $\|W_j\| = 1$ by L_2 normalization. During the testing phase, the matching score of a pair of pedestrian images is computed based on cosine similarity between the two feature vectors. This indicates that the norm of the feature vector v does not contribute to the score function. Thus, we fix $\|v\| = t$ in the training stage. Therefore, the posterior probability only depends on the cosine of the angle. To obtain a large margin classifier, we set decision boundary as follows:

$$\begin{cases} C_1 : \cos\theta_1 \geq m + \cos\theta_2 \\ C_2 : \cos\theta_2 \geq m + \cos\theta_1 \end{cases} \tag{4}$$

where $m \geq 0$ is a fixed margin parameter and it is used to better control the boundary between different classes. In Equation (4), $\cos\theta_i - m$ is smaller than $\cos\theta_i$, so that the constraint are more stringent for classification. Eventually, the modified loss enhances the discrimination of multiple granularity features by introducing an extra margin in the cosine space.

As shown in Figure 4, compared with the traditional softmax loss, there is an obvious decision boundary in large margin cosine loss. Moreover, the classification results only depend on the angle.

(**a**) Softmax loss (**b**) Large margin cosine loss

Figure 4. The decision boundary of (**a**) softmax loss and (**b**) large margin cosine loss.

Formally, the LMCL function is defined as follows:

$$Loss_{lmcl} = \frac{1}{N}\sum_{i} -\log\frac{e^{t(\cos(\theta_{l_i,i})-m)}}{e^{t(\cos(\theta_{l_i,i})-m)} + \sum_{j\neq l_i} e^{t\cos(\theta_{l_i,i})}} \tag{5}$$

In the SMGN model, the LMCL function is followed by two local branches (i.e., "Local-1" and "Local-2") and two global branches (i.e., "Global-1" and "Global-2"). Thus, we can obtain four LMCL functions, which are recorded as $Loss^1_{lmcl}$, $Loss^2_{lmcl}$, $Loss^3_{lmcl}$ and $Loss^4_{lmcl}$. Finally, we add these four LMCL functions to obtain the final identification loss function as follows:

$$Loss_{identification} = Loss^1_{lmcl} + Loss^2_{lmcl} + Loss^3_{lmcl} + Loss^4_{lmcl} \tag{6}$$

3.3.2. Verification Loss

Most previous person re-identification methods based on Siamese network regard verification process as a binary classification problem [9,27,36]. Following this idea, we adopt the widely-used cross-entropy loss function to directly compute the similarity between the extracted multiple granularity features in verification channel. For the feature pair (F_1, F_2), we compute the squared Euclidean distance as a novel feature vector in verification channel. Then the convolutional layer take the new vector as input, which is followed by a softmax output function. As a result, we can obtain a 2-dim vector (p_1, p_2) that represents the predicted probability that the two pedestrian images belong to the same person. Finally, cross-entropy loss function is formulated as follows:

$$p^s = \text{softmax}_{verification}\left(\left(F_1 - F_2\right)^2 \circ \theta_{verif}\right) \tag{7}$$

$$Loss_{verification}(\theta_{verif}, s) = \sum_{i=1}^{2} p_i^s \left(\frac{1}{p_i}\right) \tag{8}$$

where s represent the target class(same/different), $\circ$ denotes a convolutional operation, p^s is the similarity score of F_1 and F_2, and the transformation is parameterized by θ_{verif}. If the predicted result indicates that the input pedestrian image pair belongs to the same person, $p_1 = 1$, $p_2 = 0$; otherwise, $p_1 = 0$, $p_2 = 1$.

3.3.3. Fusion Loss

In order to combine the advantages of verification model and identification model, two different kind of losses mentioned above are weighted fused together to formulate the MCWF loss function as follows:

$$Loss_{fusion}(\theta, s) = \lambda Loss_{verification} + Loss_{identification} \tag{9}$$

where λ is a coefficient to balance the weight of identification and verification loss function. During the training processing, the SMGN model can guarantee multiple granularities features have the

characteristic of margin maximization for classification under the constraint of the MCWF loss function. Therefore, this type of multiple granularity features extracted from the SMGN model are regarded as large margin multiple granularities (LMMG) features. As a result, the SMGN model can improve the performance of person re-identification.

3.4. Person re-Identification Based on SMGN

During the training processing of SMGN model, given a training image set with their labels $X_{train} = \left\{ (x_t, l_t) \middle| t = 1, \ldots, N \right\}$, we first construct these images into many image pairs that are recorded as:

$$Pair = \{B_i | i = 1, 2, \ldots, T\} \tag{10}$$

where $B_i = (x_i^1, x_i^2, l_i^1, l_i^2, R_i)$ denotes the i-th image pair, R_i is the label that denotes whether x_i^1 and x_i^2 belong to the same person, if x_i^1 and x_i^2 represent the same person, $R_i = 1$; otherwise, $R_i = 0$. Based on the MCWF loss function and back propagation algorithm, the backbone CNN in SMGN model can be trained with *Pair*, which is recorded as Ω.

In the testing stage, given a query image x_q, its LMMG features F_q can be extracted by the backbone CNN Ω. Similarly, the LMMG features F_i^g of each gallery image in $X_{gal} = (x_1^g, x_2^g, \ldots, x_M^g)$ is also extracted by Ω. We compute the cosine distance between F_q and F_i^g as follows:

$$dist(F_q, F_i^g) = \frac{F_q \cdot F_i^g}{||F_q|| ||F_i^g||} = \frac{\sum_i^n F_q \times F_i^g}{\sqrt{\sum_i^n (F_q)^2} \times \sqrt{\sum_i^n (F_i^g)^2}} \tag{11}$$

where n denotes the dimension of LMMG features.

After calculating the distances between the query image x_q and each gallery image in X_{gal}, we sort these distances in ascending order to get the final ranking result. Therefore, we can calculate the corresponding right matching rates. Finally, the person re-identification procedure based on the SMGN model is summarized in Algorithm S1 (in Supplementary Materials).

4. Experiment Results

In this section, we first introduce four large-scale person re-identification databases, i.e., CUHK01, CUHK03, Market-1501 and DukeMTMC-reID. Then some experimental details are depicted, followed by some comparison with the-state-of-the-art methods on four databases. Finally, we explore the effect of the margin parameter m and the balance coefficient λ.

4.1. Datasets and Protocols

For the purpose of validating the effectiveness of the proposed model, we perform extensive experiments on four benchmark person re-identification datasets.

4.1.1. CUHK01

CUHK01 dataset is constructed by 3884 pedestrian images of 971 identities, and each identity has four images that captured by two surveillance cameras. These cameras mainly capture the front, back, left and right appearances of pedestrians. The dataset is spilt into two parts, in which 485 pedestrians are randomly selected for training and the other for testing.

4.1.2. CUHK03

CUHK03 contains 1360 people and 13,164 images captured by five non-overlapping camera pairs. Each identity is observed by two non-overlapping views and has 4.8 images under each camera on average. This dataset has two types of annotations: detector-detected (Deformable Part Model (DPM))

pedestrian bounding boxes (detected) and hand-labeled bounding boxes (labeled). All pedestrian images suffer from illumination changes, misalignment, occlusions and body part missing.

4.1.3. Market-1501

Market-1501 contains 32,668 pedestrian images of 1501 identities captured by six cameras in Tsinghua University campus. Compared with CUHK03, Market-1501 is a large scale dataset for person re-identification. In Market-1501 dataset, there are 12,396 images of 751 identities for training and 19,732 images of 750 identities for testing. All person images are detected by DPM, so some pedestrian images in Market-1501 dataset exists detection errors.

4.1.4. DukeMTMC-REID

DukeMTMC-reID is a subset of DukeMTMC that is used for multi-target tracking dataset. DukeMTMC-reID is a large scale person re-identification dataset that contains 36,411 pedestrian images of 1812 identities. The images in DukeMTMC-reID consist of 16,522 training images (from 702 people), 2228 query images (from another 702 people), and a test gallery for 17,661 images, which are captured at the Duke University campus and cropped from hand-drawn bounding boxes. The size of the images is randomly cropped, and many pedestrians are blocked.

The detail information about these datasets are summarized in Table 2. These four widely-used person re-identification datasets contain many challenges, such as misalignment, occlusions and missing body parts, low resolutions, viewpoints and background clusters. In addition, Figure 5 shows some image samples of the four datasets.

Table 2. The details of person re-identification dataset.

Dataset	Release Time	Identities	Cameras	Images	Label Method	Crop Size
CUHK01	2012	971	2	3884	Hand	160 × 60
CUHK03	2014	1467	10 (5 pairs)	13,164	Hand/DPM	Vary
Market-1501	2015	1501	6	32,217	Hand/DPM	128 × 64
DukeMTMC-reID	2017	1812	8	36,411	Hand	Vary

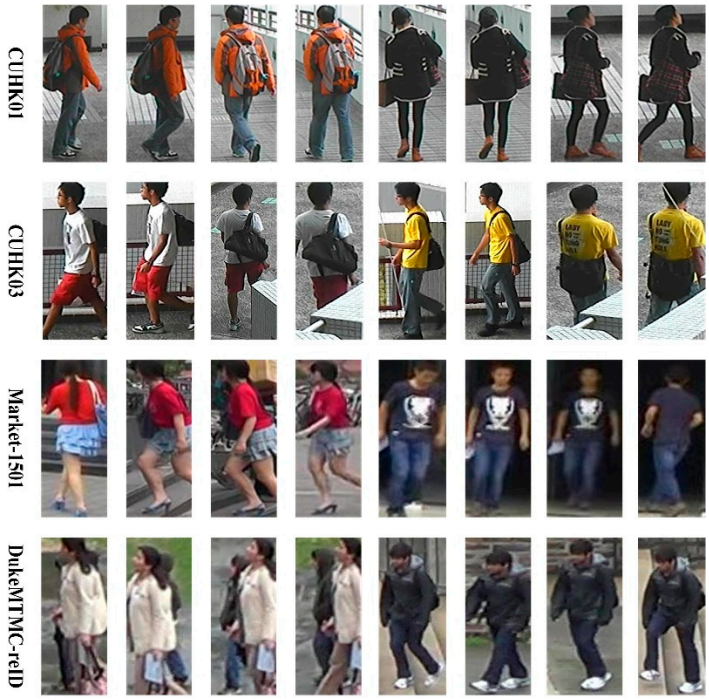


Figure 5. Some samples from CUHK01, CUHK03, Market-1501 and DukeMTMC-reID. Here, each row includes two different identities captured under different cameras.

4.1.5. Metric Protocols

As an evaluation protocol, cumulative match characteristic (CMC) is extensively applied in person re-identification to count the ranks of true matches. At the same time, we also introduce the mean average precision (mAP) for the Market-1501 and DukeMTMC-reID datasets in our experiment. These two criteria are executed under a single query setting for the four datasets. More importantly, the re-ranking method based on the k-reciprocal encoding [37] is adopted for further improvement.

4.2. Implementation Details

We use Python to implement the proposed SMGN model. Some details about data preparation, parameter settings and data augmentation are described in this section.

4.2.1. Data Preparation

For the convenience to extracting features of pedestrian images, we perform the input data preparation. Firstly, we resize all the images into 256×256. Then we utilize the resized input images to subtract the mean image. Afterwards, a random order style [11] is introduced in our paper and we set the initial ratio of positive images to negative images to improve the performance of the SMGN model. In the end, we multiple the ratio between positive and negative pairs by a factor of 1.01 every epoch until it reaches 3:1 to prevent our model from over-fitting.

4.2.2. Parameter Settings

In this experiment, we set the size of image batch to 32 for SMGN, including eight positive and eight negative image pairs. Stochastic gradient descent (SGD) is adopted to update the parameter of SMGN model. The number of training epoch is set to 1000. We set the weight decay to 5×10^{-4} and the momentum to 0.9. As for the learning rate, we set the initial learning rate to 0.001 and then set to 0.0001 for the last 10 epochs. When perform binary-class task, we randomly select negative pairs from the whole negative sample pool for each batch. For the network updating, we accumulate all the gradients produced by every image pair. In the training phase, the weight of the gradient generated by the verification loss is three times as much as the identification loss. We set the parameters $\lambda = 3$ and $m = 0.40$ empirically in all the following experiments. The validation experiment as Figures 6 and 7 illustrated.

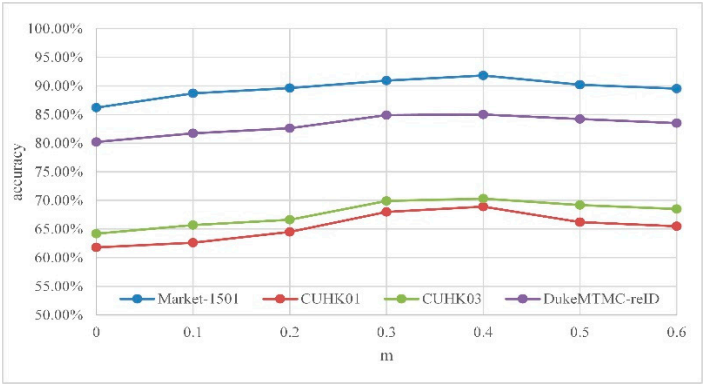


Figure 6. Rank-1 performance of SMGN with different margin parameter m on Market-1501, CUHK01, CUHK03 and DukeMTMC-reID.

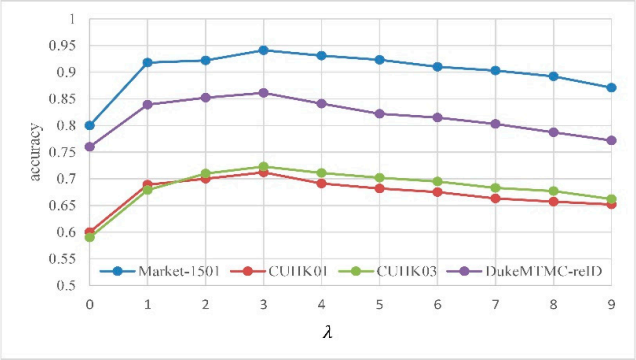


Figure 7. Rank-1 performance of SMGN with different coefficient λ on Market-1501, CUHK01, CUHK03 and DukeMTMC-reID.

4.2.3. Data Augmentation

Person re-identification datasets are composed by various images of different pedestrians, in which each pedestrian has a limited number of images. Because of this, we cannot construct adequate positive pairs to train the SMGN model. Therefore, there exists over-fitting and the performance of the Siamese network is poor.

Compared with the other datasets, CUHK01 is a small scale person re-identification dataset. To cope with over-fitting since the lack of data, data augmentation is adopted in our experiment. Specifically, all the resized pedestrian images are randomly cropped to 224×224 at first. Besides that, horizontal flipping is used on the CUHK01 dataset to implement image augmentation.

4.3. Parameter Analysis

In this section, we evaluate two important parameters, i.e., the fixed margin parameter m in Equation (5) and the balance coefficient λ in Equation (9).

4.3.1. Effect of m

The margin parameter m plays an important role in LCML. To investigate the effort of m, we conduct a comparative experiment in this part. For Figure 6, we compare the results with different margin parameter on CUHK01, CUHK03 (labeled), Market-1501 and DukeMTMC-reID. The margin parameter is used to better control the boundary between different classes. If the margin rate is too large, then the model will fail to converge. In this part, we set the range of m as $[0, 0.6]$ and for every 0.1 m increase, we do a comparison experiment once more. As shown in Figure 6, we can find that the matching performance is worst when $m = 0$ on the four person re-identification datasets. As m being increased, the accuracy of the proposed model in every dataset consistently improves and get saturated at $m = 0.40$. For convenience, the parameter m in Equation (6) is set to fixed 0.40 in the subsequent experiments. Note that λ is set to 1 in this part.

4.3.2. Effect of λ

The balance coefficient λ is to balance verification loss and identification loss. To investigate the effort of λ, we conduct a comparative experiment as Figure 7 illustrated (Note that m is set to 0.40). In this part, we set the range of λ as $[1, 9]$ and for every 1 m increase, we do a comparison experiment once more. As shown in Figure 7, we can see that the matching rates are lowest on the four datasets when $\lambda = 0$. In other words, we cannot obtain the best performance if we only use identification model. Because the identification model only makes full use of the label information of pedestrian images, which is benefit to intra-class separation. As for inter-class compactness, we assume that the verification loss equals zero if the two images belong to the same identity. So we can see that the matching degree is higher with the increase of weight coefficient λ. When λ is set to 3, we can get

the good performance on CUHK01, CUHK03 (labeled), Market-1501 and DukeMTMC-reID. In the following experiment, the parameter λ in Equation (7) is set fixed to 3 in this paper.

4.4. Performance Evaluation

We compare the proposed SMGN model with current state-of-the-art approaches on the four widely-used datasets to show our competitive performance. Comparative results in detail are given below.

4.4.1. Performance on the CUHK01 Dataset

Compared with the state-of-the-art results reported on the CUHK01 dataset, the proposed SMGN model show the best performance that are listed in Table 3. For CUHK01, we consider 486 identities for testing and the rest for training like most previous papers. As shown in Table 3, we can observe that the proposed SMGN model achieve the best rank-1 matching rate at 71.2%, which is higher 2.1% higher than the second best one NFST [38]. With the re-ranking technique in [37], we obtain a higher rank-1 rate on CUHK01.

Table 3. Comparison with the several results reported on the CUHK01 dataset using a CMC curve.

Method	Rank 1	Rank 5	Rank 10	Rank 20
FPNN [9]	27.9%	—	—	—
Deep CNN [10]	47.5%	—	—	—
KCVDCA [39]	47.8%	74.2%	83.4%	89.9%
LOMO+XQDA [4]	49.2%	75.7%	84.2%	90.8%
TCP [40]	53.7%	84.3%	91.0%	93.3%
GOG+XQDA [6]	57.8%	79.1%	86.2%	92.1%
NFST [38]	69.1%	86.9%	91.8%	95.4%
Ours	**71.2%**	**87.2%**	**90.9%**	**95.5%**
Ours+re−rank	**72.0%**	**88.1%**	**91.2%**	**96.3%**

4.4.2. Performance on the CUHK03 Dataset

The CUHK03 dataset has two types of annotations as mentioned above, i.e., labeled and detected. As we can see that the results using different methods on CUHK03 are shown in Table 4. We have the same settings as [9], that is, CUHK03 is partitioned into a training set (1160 persons), validation set (100 persons), and test set (100 persons). It is clear that the proposed SMGN outperforms the other existing methods in the case of both detected and labeled. In Table 4, we can see that the proposed algorithm achieves 70.2% at rank 1 in the case of detected boxes and 72.3% with manual bounding boxes. With the re-ranking technique described in [37], we got a better performance in both cases.

Table 4. Comparison with the several results reported on the CUHK03 dataset using a CMC curve.

Method	Detected			Labeled		
	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10
FPNN [9]	19.9%	49.0%	64.3%	20.7%	51.7%	68.3%
DPFL [41]	40.7%	—	—	43.0%	—	—
SVDNet [42]	41.5%	—	—	40.9%	—	—
HA-CNN [43]	41.7%	—	—	44.4%	—	—
Deep CNN [10]	45.0%	75.7%	83.0%	54.7%	88.3%	93.3%
LOMO+XQDA [4]	46.3%	79.0%	88.6%	52.2%	82.3%	92.1%
MGCAM [44]	46.7%	—	—	50.1%	—	—
LOMO+MLAPG [8]	51.2%	—	—	58.0%	—	—
NFST [38]	54.7%	84.8%	94.8%	62.6%	90.1%	94.8%
PCB+RPP [45]	63.7%	80.6%	86.9%	—	—	—
GOG+XQDA [6]	65.5%	88.4%	93.7%	67.3%	91.0%	96.0%
MGN [16]	66.8%	—	—	68.0%	—	—
Ours	**70.2%**	**87.2%**	**93.9%**	**72.3%**	**89.1%**	**96.7%**
Ours+re−rank	**71.5%**	**88.3%**	**94.0%**	**73.1%**	**90.0%**	**97.1%**

4.4.3. Performance on the Market-1501dataset

We summarize the performance results on Market-1501 dataset using some state-of-the-art methods and our proposed algorithm. It can be found that the deep learning based methods (i.e., Gated SCNN [19], DPFL [42], PCB+RPP [46] etc.) obviously defeat non-deep learning based methods (i.e., BoW+kissme [28], LOMO+XQDA [4]) on the Market-1501 dataset. We can see that the proposed SMGN obtains 94.2% and 80.2% in rank-1 and mAP accuracy respectively. With the re-ranking technique [38], the proposed algorithm outperforms the second best one by a margin of 1.7% at rank-1 under the single query (SQ) setting.

4.4.4. Performance on DukeMTMC-reID

From Table 5, we can see that our algorithm on the DukeMTMC-reID dataset achieves 87.1% rank-1 matching rate and 76.0% mAP respectively, which significantly outperforms the previous state-of-the-art methods. The results on the DukeMTMC-reID dataset show that our method has a great advantage on large scale dataset. Compared with the state-of-the-art methods, our proposed method obtains competitive results on all four datasets. Especially, SMGN achieves 71.2% rank-1 accuracy for CUHK01, 70.2% rank-1 accuracy for CUHK03 (detected), 72.3% rank-1 accuracy for CUHK03 (labeled), 94.1% for Market-1501 and 86.1% for DukeMTMC-ReID without re-ranking. In addition, we visualize the top-10 ranking results on Market-1501 for some randomly-selected query pedestrian images in Figure 8. The results indicate the good performance of our model.

Table 5. Comparison with the several results reported on the Market-1501 and DukeMTMC-reID datasets using a CMC curve.

Method	Market-1501		DukeMTMC-re-ID	
	Rank-1	MAP	Rank-1	MAP
BoW+kissme [18]	39.6%	17.7%	25.1%	12.2%
LOMO+XQDA [4]	43.8%	22.2%	30.8%	17.0%
NFST [38]	55.4%	29.9%	—	—
Gated SCNN [27]	65.9%	39.6%	—	—
SVDNet [42]	82.3%	62.1%	76.7%	56.8%
MGCAM [44]	83.8%	74.3%	—	—
PSE [46]	87.7%	69.0%	79.8%	62.0%
DPFL [41]	88.6%	72.6%	79.2%	60.6%
HA-CNN [43]	91.2%	75.7%	80.5%	63.8%
Deep-Person [47]	92.3%	79.6%	80.9%	64.8%
PCB+RPP [45]	93.8%	81.6%	83.3%	69.2%
Ours	**94.1%**	**79.2%**	**86.1%**	**75.3%**
Ours+re-rank	**95.5%**	**80.3%**	**87.1%**	**76.0%**

Figure 8. Three example query images in Market-1501 test set and their corresponding top 10 ranking lists results using our method. The green boundary means true positive and red means false positive.

5. Conclusions

In this paper, we propose a novel symmetric Siamese model named SMGN for person re-identification. In order to learn multiple granularity features from global and local regions, we adopt modified ResNet-50 as the backbone network at first and use the local and global branches to extract multiple granularity features. Then a multi-channel weighted fusion (MCWF) loss function is designed to further reduce the intra-class variance while increase the inter-class variance, which consider an obvious decision boundary when classifying. Finally, we integrated SMGN and the MCWF loss function together and the large margin multiple granularities (LMMG) features can be obtained when the loss function tends to the minimum value. After waiting for SMGN to stabilize, we use the backbone network of it for testing to get the ranking lists of the target image. We validated the effectiveness of the proposed SMGN on four widely-used person re-identification datasets and the performance on those are improved comparing with many state-of-the-art methods. Our future work is to explore more robust and discriminative features of person images and investigate on how to achieve compactness of intra-class and separation of inter-class much better.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-8994/12/1/92/s1.

Author Contributions: Supervision: S.-W.T.; validation: G.-Y.F.; Writing—original draft, G.-Y.F.; writing—review and editing: D.-X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the International Cooperation and Exchange Foundation of Shaanxi, China (Grant 2017KW-013), and the Innovation & Entrepreneurship Dual Tutor Foundation of Shaanxi, China (grant nos. 2019JM-604), and the Xi'an University of Posts and Telecommunications Graduate Innovation Fund Project under grant CXJJLY2018040.

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

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-Identification: Past, Present and Future. Available online: https://arxiv.org/abs/1610.02984 (accessed on 5 June 2019).
2. Karanam, S.; Gou, M.; Wu, Z.; Rates-Borras, A.; Camps, O.; Radke, R.J. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 523–536. [CrossRef] [PubMed]
3. Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; Cristani, M. Person re-identification by symmetry-driven accumulation of local features. In Proceedings of the Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2360–2367.
4. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
5. Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; Li, S. Salient color names for person re-identification. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Volume 8689, pp. 536–551.
6. Matsukawa, T.; Okabe, T.; Suzuki, E.; Sato, Y. Hierarchical Gaussian descriptor for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1363–1372.
7. Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
8. Liao, S.; Li, S. Efficient PSD constrained asymmetric metric learning for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Santiago, Chile, 11–18 December 2015; pp. 3685–3693.
9. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep filter pairing neural network for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.

10. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
11. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned CNN embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *14*, 1–20. [CrossRef]
12. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint detection and identification feature learning for person search. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3376–3385.
13. Wu, L.; Wang, Y.; Gao, J.; Li, X. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Trans. Multimed.* **2019**, *21*, 1412–1424. [CrossRef]
14. Chung, D.; Tahboub, K.; Delp, E.J. A two stream siamese convolutional neural network for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 1983–1991.
15. Yan, Y.; Ni, B.; Song, Z.; Ma, C.; Yan, Y.; Yang, X. Person re-identification via recurrent feature aggregation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 701–716.
16. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning Disriminative Features with Multiple Granularities for Person Re-Identification. Available online: https://arxiv.org/abs/1804.01438 (accessed on 25 June 2019).
17. Li, W.; Zhao, R.; Wang, X. Human re-identification with transferred metric learning. In *Lecture Notes in Computer Science, Proceedings of the Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 31–44.
18. 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, 11–18 December 2015; pp. 1116–1124.
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. 3754–3762.
20. He, N.; Paoletti, M.E.; Haut, J.M.; Fang, L.; Li, S.; Plaza, A.J.; Plaza, J. Feature extraction with multiscale covariance maps for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 755–769. [CrossRef]
21. Shih, Y.; Yeh, Y.; Lin, Y.; Weng, M.; Lu, Y.; Chuang, Y. Deep co-occurrence feature learning for visual object recognition. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7302–7311.
22. Gupta, U.; Chatterjee, A.; Srikanth, R.; Agrawal, P. A Sentiment-and-Semantics-Based Approach for Emotion Detection in Textual Conversations. July 2017. Available online: https://arxiv.org/abs/1707.06996 (accessed on 5 August 2019).
23. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 384–393.
24. 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.
25. Xiao, T.; Li, H.; Ouyang, W.; Wang, X. Learning deep feature representations with domain guided dropout for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1249–1258.
26. Subramaniam, A.; Chatterjee, M.; Mittal, A. Deep neural networks with inexact matching for person re-identification. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2667–2675.
27. Varior, R.R.; Haloi, M.; Wang, G. Gated siamese convolutional neural network architecture for human re-identification. In Proceedings of the European Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 791–808.

28. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 499–515.
29. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. Available online: https://arxiv.org/abs/1703.07737 (accessed on 1 October 2019).
30. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-Margin Softmax Loss for Convolutional Neural Networks. Available online: https://arxiv.org/abs/1612.02295 (accessed on 10 October 2019).
31. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. SphereFace: Deep hypersphere embedding for face recognition. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6738–6746.
32. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive margin softmax for face verification. *Signal Process.* **2018**, *25*, 926–930. [CrossRef]
33. Almazán, J.; Gajic, B.; Murray, N.; Larlus, D. Re-Id Done Right: Towards Good Practices for Person Re-Identification. Available online: https://arxiv.org/abs/1801.05339 (accessed on 13 December 2019).
34. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lile, France, 6–11 July 2015; pp. 448–456.
35. 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 Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5265–5274.
36. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 December 2014; pp. 34–39.
37. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3652–3661.
38. Zhang, L.; Xiang, T.; Gong, S. Learning a discriminative null space for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1239–1248.
39. Chen, Y.; Zheng, W.; Lai, J.; Yuen, P. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1661–1675. [CrossRef]
40. 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 Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1335–1344.
41. Chen, Y.; Zhu, X.; Gong, S. Person re-identification by deep learning multi-scale representations. In Proceedings of the Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 2590–2600.
42. 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. 3800–3808.
43. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2285–2294.
44. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model for personre-Identification. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1179–1188.
45. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision*; Springer: Munich, Germany, 8–14 September 2018; pp. 480–496.
46. 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 Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 420–429.
47. Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; Xu, Y. Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. Available online: https:/arxiv.org/abs/1711.10658 (accessed on 10 November 2019).