*algorithms*

MDPI

*Article*

# Multi-Level Joint Feature Learning for Person Re-Identification

**Shaojun Wu [1,2] and Ling Gao [1,2,*]**

[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong Province, China; wusj98287@gmail.com

[2] Institute of Data Science and Technology, Shandong Normal University, Jinan 250014, Shandong Province, China

* Correspondence: gaoling@sdnu.edu.cn

Received: 1 April 2020; Accepted: 27 April 2020; Published: 29 April 2020


**Abstract:** In person re-identification, extracting image features is an important step when retrieving pedestrian images. Most of the current methods only extract global features or local features of pedestrian images. Some inconspicuous details are easily ignored when learning image features, which is not efficient or robust to for scenarios with large differences. In this paper, we propose a Multi-level Feature Fusion model that combines both global features and local features of images through deep learning networks to generate more discriminative pedestrian descriptors. Specifically, we extract local features from different depths of network by the Part-based Multi-level Net to fuse low-to-high level local features of pedestrian images. Global-Local Branches are used to extract the local features and global features at the highest level. The experiments have proved that our deep learning model based on multi-level feature fusion works well in person re-identification. The overall results outperform the state of the art with considerable margins on three widely-used datasets. For instance, we achieve 96% Rank-1 accuracy on the Market-1501 dataset and 76.1% mAP on the DukeMTMC-reID dataset, outperforming the existing works by a large margin (more than 6%).

**Keywords:** deep learning; intelligent monitoring; person re-identification

## 1. Introduction

Public safety incidents often occur in dense crowds. Therefore, a large number of surveillance cameras are installed and applied in various areas of the city. Person re-identification is a key component technology in the field of urban remote sensor monitoring. For a certain target person appearing in a remote sensing surveillance video or remote sensing pedestrian image, the method of person re-identification can accurately and quickly identify this target person in other remote sensing monitoring fields. The goal of person re-identification is to find the same person from the videos or images captured from different cameras [1], as in Figure 1. Recently, deep learning methods achieve great success by designing feature representations [2–6] or learning robust distance metrics [7–10].

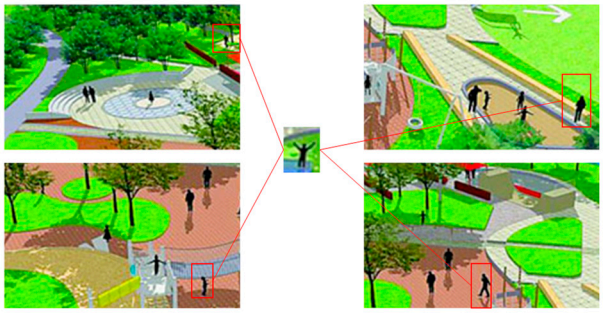


**Figure 1.** Retrieving the same pedestrian image under different cameras.

The pedestrian features extracted by deep learning can be divided into two types: global features and local features. The global features are extracted from the whole picture, which is easy to calculate and intuitive. These features contain the most significant information of a person (such as the color of pedestrian clothes), which is helpful to indicate the identity of different pedestrians [6]. However, some inconspicuous details (such as hats, belts, etc.) are easily ignored by the global features. For example, if two persons are wearing clothes of the same color, and one of them is wearing a hat, it is hard to discriminate the two persons from only the overall appearance. Moreover, when the background is complex, it is difficult for the global features to associate the images of the same person with different backgrounds into one identity, as shown in Figure 2.

In order to solve the problem of person re-identification, some recent work mainly uses deep learning models to extract local features, using salient local details to match the local features of a queried pedestrian. Local feature information of each body part is extracted by neural network. The similarity between local features is very low, which is more conducive to person re-identification. However, the method of extracting local features may ignore the overall pedestrian information, as shown in Figure 2.

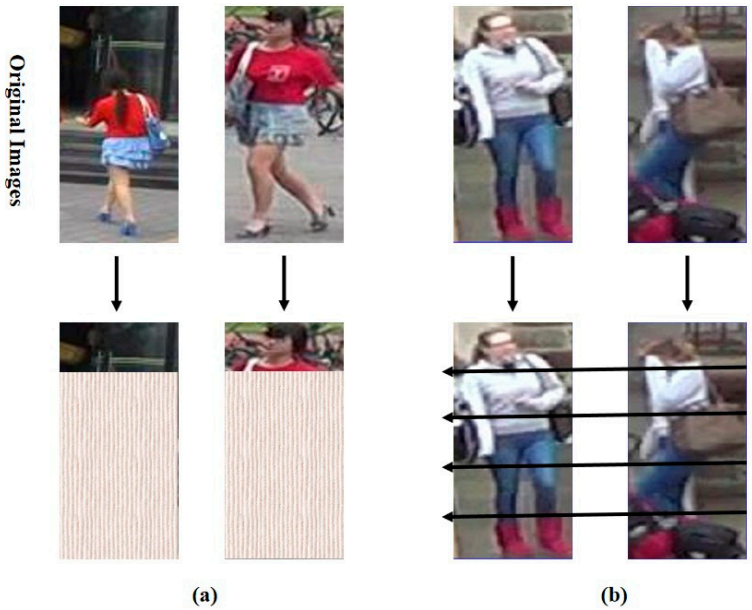



**Figure 2.** If the network only extracts local features of (**a**), it cannot be determined that those local features belong to the same person. If the network only extracts global features, complex background content can be detrimental to identifying pedestrians, as shown in (**b**). Horizontal arrows mean we horizontally divided the feature map into six parts and extracted the local features of each part.

Representing local information of individuals by locating notable body parts from pedestrian images is also an effective method of person re-identification in recent years [11–13]. Local features are

extracted from different body parts. Each body part contains a small portion of local information from the whole body [14,15]. In this way, we can learn detailed local features from the divided parts which make the features focus on the local details. The learned local features supplement important details, which can be taken as the complementary of the global features.

Therefore, in this paper, the local features and global features are jointly learned for person re-identification. In this paper, we propose a Multi-level Feature Fusion (MFF) model that fuses global features and local features. Moreover, the local features are extracted from different network depths. An MFF model consists of two components: Part-based Multi-level Net (PMN) and Global-Local Branch (GLB). PMN is used to extract local features from different layers of the network. GLB extracts local features and global features at the highest level. The global features and local features are used to perform identity predictions in MFF. We train the MFF model on three classic datasets. Performance of experiments show that the MFF model which fuses global and local features is particularly effective and our model results outweigh many state-of-the-art methods.

The main contributions of our work are as follows:

- We add Part-based Multi-level Net (PMN) to extract local features more comprehensively from lower to higher layers of the network. Compared with other traditional feature extraction methods, PMN can learn more local detailed features from different network layers.
- We join to learn global features and local features. The robustness of the MFF model can be improved by joint learning features. We use multi-class loss functions to classify the features extracted from different network branches separately, which enhances the accuracy of MFF.
- We implement extensive experiments on three challenging person re-identification datasets. Experiments show that our method is superior to existing person re-identification 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 three benchmark datasets are shown in Section 4. The conclusions of our work are described in Section 5.

## 2. Related Work

Person re-identification aims to find matching pedestrian images from different camera views. With the rapid development of deep learning, feature learning by deep networks has become a common practice in person re-identification. Li et al. [16] combined deep siamese network architecture with pedestrian human body feature learning for the first time and achieved higher performance. Zheng et al. [11] proposed a baseline that combined ID-discriminative embedding (IDE) with a ResNet-50 backbone network for modern deep person re-identification systems. Proposed methods also improved the performance of deep person re-identification. Varior et al. [17] described the interrelationship of local parts by computing mid-level features of image pairs. Xiao et al. [18] improved the generalization of different pedestrian scenes by using the Domain Guided Dropout method. Yang et al. [19] used deep learning networks to integrate multiple feature representations together for person re-identification. Some recent works use features of different views or top-view for person re-identification [20,21]. Paolanti et al. [21] extracted neighborhood component features and used multiple nearest neighbor classifiers to identify pedestrians.

For local features extraction, Li et al. [12] proposed a deep learning method called STN. Local features can be easily localized from image patches by learning deep contextual awareness of body and potential parts. Zhao et al. [13] applied deep learning method to align same parts of different images after splitting pedestrian image. Liu et al. [22] utilized attention module to extract part features emphasized of the model. Bai et al. [23] combined some feature slices which are vertically divided into multiple pieces with the LSTM network. Some recent works strengthen the representation of the body part by embedding attention information [22,24,25]. In our proposed method, we extract local features

from several horizontal stripes. At the same time, local features are extracted from different network depths which achieves good performance.

For global feature extraction, a kernel feature map is used to obtain similar information of all patches from different images [26]. Liao et al. [6] proposed a method called Local Maximal Occurrence (LOMO) to represent a local feature which has a positive effect on person re-identification. In our paper, we combine global features and low-to-high level local features together for person re-identification.

In the feature learning phase, classification loss is a commonly used loss function. Some loss functions based on softmax loss achieve state-of-the-art performance in face recognition. Liu et al. [27] proposed L-Softmax to improve the discrimination of pedestrian image features by adding angular constraints to each identity. A-Softmax [28] improves L-Softmax by normalizing weights to recognize by learning angularly discriminative features. Since softmax loss is robust to various multi-class classification tasks and can be used individually [25,29] or in combination with other losses [10,16,23,30–32], softmax loss is often used as a classification for loss function in person re-identification. In our proposed method, we also use softmax loss to solve multi-class tasks.

## 3. Materials and Methods

Details of our method are described in this section. Proposed Multi-level Feature Fusion (MFF) is introduced in detail which contains two main components: Part-based Multi-level Net (PMN) and Global- Local Branch (GLB), as shown in Figure 3. PMN is used to extract local features from different layers of the network. GLB extracts local features and global features from the final layer. More details of the MFF are introduced in Sections 3.1 and 3.2. The loss function is introduced in Section 3.3.

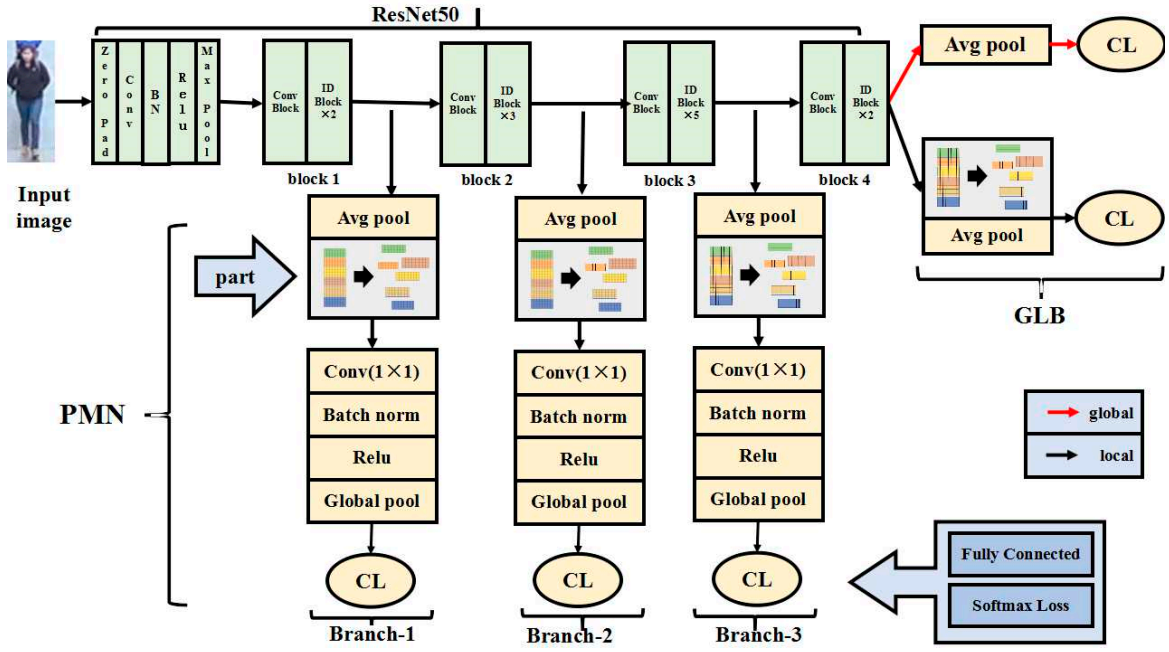



**Figure 3.** Multi-level Feature Fusion (MFF) architecture is split into two components: Multi-level Net (PMN) and Global-Local Branch (GLB). An input image is fed to the modified ResNet50 to obtain feature maps. The extracted global features are sent to a classifier. Meanwhile, the feature maps are divided into several parts and each part after dimension reduction is sent to a classifier.

### *3.1. Change of Backbone Network*

By considering the relatively effective performance and concise architecture, this paper uses ResNet50 as the backbone network. In order to extract features more accurately, the ResNet50 structure is divided into block1, block2, block3 and block4, as in Figure 3. Then we can extract feature maps between each block and use classifiers to predict identity. Each block consists of conv block (includes

multiple convolution blocks) and identity blocks. The upper layer network of block1 is max pooling layer. The backbone structure of ResNet50 remains unchanged until block4. In this paper, we remove the entire network after block4 (including the global average pooling layer). In this way, feature maps will be retained with more feature information by removing the global average pooling layer.

### *3.2. Structure of Multi-Level Feature Fusion (MFF)*

The combination of global features and local features can learn more information which leads to more accurate pedestrian retrieval results. In this paper, we propose an MFF model which fuses local features and global features together. In MFF, local features and global features are learned identity predictions. As shown in Figure 3, the MFF model is composed of Part-based Multi-level Net (PMN) and Global-Local Branch (GLB). The structure of PMN and GLB are introduced separately. And Table 1 shows the dimensions of the features extracted from each branch.

**Table 1.** Comparison of the settings for five branches.

| Branch | Dimension |
|---|---|
| Branch-1 | $256 \times 6$ |
| Branch-2 | $256 \times 6$ |
| Branch-3 | $256 \times 6$ |
| GLB-1 | 256 |
| GLB-2 | $256 \times 6$ |

The structure of Global-Local Branch (GLB) consists of two parts to extract local and global features from the deepest layer of the network, respectively. Given an input image, we can obtain the feature maps through the backbone network. Then an average pooling layer and a classifier are employed after the ResNet50 network to get the 256-dimension global features. The classifier is composed of a fully connected layer and a softmax layer to get the prediction of pedestrian identity from the global feature. The second branch of GLB is used to extract local features from the deepest layer of the network. In order to extract the local features, we divide the feature maps horizontally into six parts as shown in Figure 3. We add an average pooling layer and a classifier after the divided feature map to get the prediction of pedestrian identity.

The structure of Part-based Multi-level Net (PMN) consists of three parts (Branch-1, Branch-2 and Branch-3) which is used to extract local features from lower to higher layers of the network, as shown in Figure 4. ResNet50 consists of four blocks, and we add Branch-1, Branch-2 and Branch-3 between each pair of continuous blocks. In each branch, firstly, we apply an average pooling on the corresponding output feature map. Then the feature map is divided into six parts horizontally as introduced in previous subsection. We add a $1 \times 1$ kernel-sized convolutional layer, a batch normalization layer, a relu layer and a global pooling layer to obtain $6 \times 256$-dimension local features. Then each local feature is input into a classifier, where each classifier is implemented with a fully-connected (FC) layer and a softmax layer. The classifier is used for the identity prediction. Note that, Branch-1, Branch-2 and Branch-3 run in parallel.


**Figure 4.** Method of body part partitions is shown above. The feature maps extracted from the deep learning model is horizontally divided into six parts.

### *3.3. Loss Function*

In our paper, we regard the person identification task as a multi-class classification problem. Considering that softmax loss is widely used in various deep person re-identification methods, we employ softmax loss as the loss function for classification in training stage.

In MFF, we regard person re-identification task as a multi-class classification problem. For *i*-th learned class vector $h_i$, the softmax loss function is described as follows:

$$L_{softmax} = -\sum_{i=1}^{M} \log \frac{\exp\left(K_{yi}^{T} h_i\right)}{\sum_{c=1}^{D} \exp\left(K_c^{T} h_i\right)} \tag{1}$$

where $K_c$ is the weight of class $c$, $D$ is the number of classes in training dataset, $K_{yi}$ is the weight of $yi-th$ in fully connected layer, $yi-th$ is the $i-th$ value of output vector $y$. $M$ is the size of mini-batch in training process. In MFF, the softmax loss is employed into the features extracted by GLB and PMN.

The final loss function is formulated as follows:

$$L = L_{softmax}^{G} + L_{softmax}^{L} + L_{softmax}^{L_1} + L_{softmax}^{L_2} + L_{softmax}^{L_3} \tag{2}$$

where $L_{softmax}^{G}$ and $L_{softmax}^{L}$ represent the identity classification tasks in global and local branches of GLB, $L_{softmax}^{L_1}$, $L_{softmax}^{L_2}$ and $L_{softmax}^{L_3}$ represent the identity classification tasks in Branch-1, Branch-2 and Branch-3 of PMN, respectively.

Each classifier predicts the most similar pedestrian images when using a single classifier to make decisions. A pedestrian image with the same identity as a query image is usually determined as the most similar pedestrian image in the process of classification. We vote on the prediction results obtained by five classifiers to get the final classification prediction results.

## 4. Results

### *4.1. Datasets*

In order to evaluate the performance of the MFF model, here we evaluate three datasets in the experiments, i.e., Market-1501 [33], DukeMTMC-reID [6] and CUHK03 [34]. The dataset of person re-identification is divided into Training_set, Verification_set, Query and Gallery. In our experiment, the network model is trained on the training set. Then we calculate the similarity of features extracted from Query and Gallery which is used to find similar pedestrian images of Query in Gallery. Pedestrian images of the Gallery are sorted according to the similarity of image features, as shown in Figure 5.

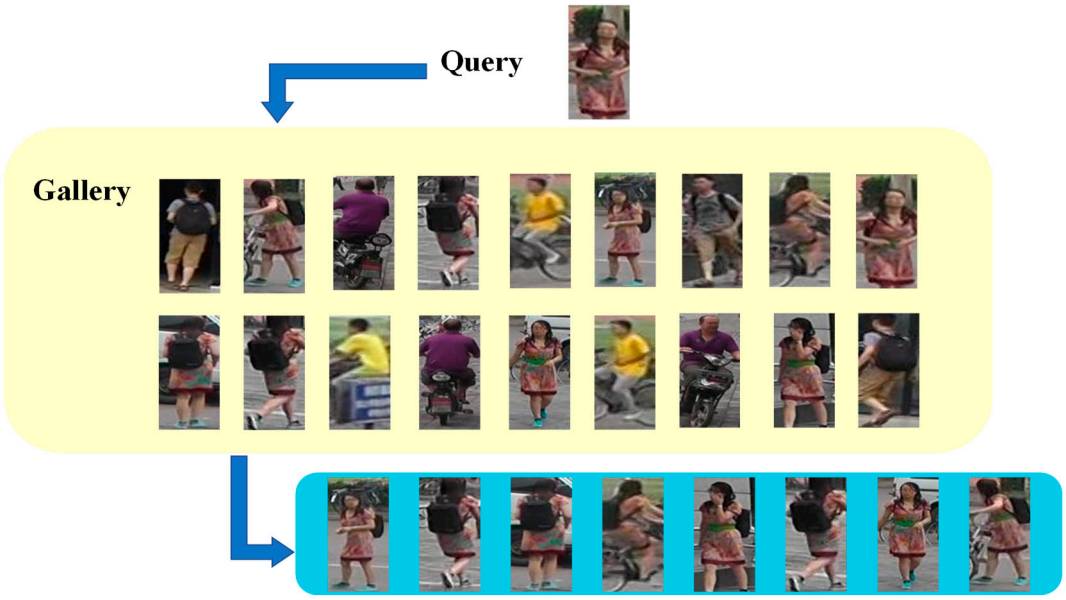



**Figure 5.** Examples of person re-identification. The similar pedestrian images of Query are shown in blue box.

The Market-1501 [33] dataset includes 1501 identities captured by six cameras and 32,668 detected pedestrian rectangles under six camera viewpoints. In this dataset, each pedestrian contains at least two camera viewpoints. The training set is consisted of 751 identities and each identity includes 17.2 training data on average. The test set is composed of 19,732 images of 750 identities. The pedestrian detection rectangle in the gallery is detected by DPM [35]. Here, we use mean Average Precision (mAP) to evaluate person re-identification algorithms.

The DukeMTMC-reID [6] dataset consists 36,411 images of 1404 identities. With those images collected by eight cameras and each image sampled every 120 frames from the video. This dataset is composed of 16,552 training images, 2228 query images and 17,661 gallery images. Half of the identities are randomly sampled as training sets while the others as test sets. DukeMTMC-reID offers human labeled bounding boxes.

The CUHK03 [34] dataset is composed of 13,614 images and 1467 identities. Each identity automatically captured by two cameras. In this dataset, bounding boxes are provided by two different ways: automatically detected which is the same as Market-1501 dataset and hand-labeled bounding boxes. Here we use two kinds of bounding boxes in this paper. In the whole experiment, we evaluate the single-query setting and adopt new test protocol proposed in [36] which is similar to Market-1501. CUHK03 is divided into a training set consisting of 756 pedestrians and a test set of 700 pedestrians in the new protocol. A randomly selected image is used as query image while the rest is used as gallery. In this way, each pedestrian gets multiple ground truths in gallery.

The detailed information about these datasets is summarized in Table 2. Three widely-used person re-identification datasets contain many challenges, such as misalignment, low resolutions, viewpoints and background clusters. In addition, Figure 6 shows some image samples of the four datasets.

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

| Dataset | Release Tine | Identities | Cameras | Crop Size | Label Method |
|---|---|---|---|---|---|
| Market-1501 | 2015 | 1501 | 6 | Vary | Hand/DPM |
| DukeMTMC-reID | 2017 | 1812 | 8 | $128 \times 64$ | Hand |
| CUHK03 | 2014 | 1467 | 10 | Vary | Hand/DPM |

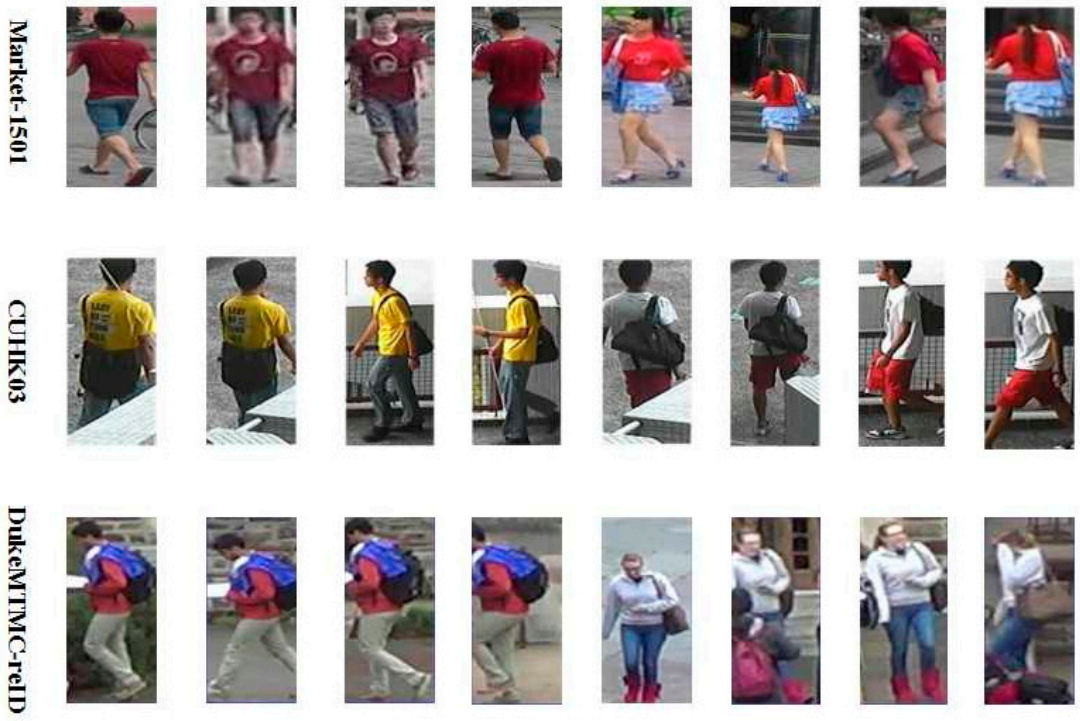



**Figure 6.** Some samples from Market-1501, CUHK03 and DukeMTMC-reID. Here each row includes two different identities captured under different cameras.

For each query image, we merge the five feature vectors into one and calculate the Euclidean distance between query image and pedestrian image in gallery. We use the Euclidean distance value to rank the images. The higher the ranking, the more similar the image is to the query image. Then we arrange them in descending order according to the Euclidean distance, and use the Cumulative Match Characteristic (CMC) curve to show the performance. In terms of performance measurement, we use the Rank-1 accuracy and the mean Average Precision (mAp).

Mean Average Precision (mAP) is an important evaluation indicator for person re-identification. Precision and recall are important components of mean Average Precision. Precision is the ability of a model to identify only the relevant objects. Recall is the ability of a model to find all the relevant cases. The precision and recall are expressed as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

where TP means the number of true positive, FP means the number of false positive, FN means the number of false negative.

Average precision (AP) means the mean of the highest precision under different recalls, which is expressed as follows:

$$AP = \frac{1}{R}\sum_{r \in R} \text{Precision}(r) \tag{5}$$

MAP is the average value of the AP, which is expressed as follows:

$$mAP = \frac{1}{M}\sum_{m \in M} AP(m) \tag{6}$$

### 4.2. Implementation Details

We pre-trained ResNet50 on ImageNet [37] and used the weight of ResNet50 in MFF. Our training environment is Pytorch and code is edited using python. The computer configuration system is 64-bit ubuntu 16.04LTS. Single-GPU training is used in MFF and the type of GPU is NVIDA GEFORCE GTX 1080. Considering the configuration of the graphics card, we set batch size to 32. Due to differences between three datasets, the learning rate of each dataset is different. Learning rate of Market-1501 is 0.05. Learning rate is set to 0.045 when training on DukeMTMC-reID. The learning rate of CUHK03 is 0.08. The entire training process is terminated in 60 epochs. We randomly select one image as the query image which means we conduct all the experiments under single-query settings, and the input pedestrian images are resized to $384 \times 192$.

### 4.3. Comparison with Market1501

Comparison with the proposed method on Market-1501 is detailed in Table 3. The MFF model is compared with several state-of-the-art person re-identification methods on Market-1501 in recent years, for example, the bag of words model BoW+KISSME [33] with a hand-crafted method, the SVDNet [34] using global features extracted by deep learning model, and the part-aligned representation PAR [17] using part features extracted by a deep learning model. We can observe from Table 3 that the proposed MFF model gets best results in Rank-1 accuracy, Rank-5 accuracy and Rank-10 accuracy. In the experiment, we use mean average precision (mAP) as an evaluation index of person re-identification. The MFF model achieves 87.9% mAP on the Market-1501, which is 18.8% higher than the best proposed method. In addition, the MFF model achieves Rank-1 accuracy of 96.0%, which is 11.1% higher than the best method. Rank-5 accuracy of our model achieves 98.7%, 4.5% better than the best compared method. This is because the MFF model fuses the global features and local features together. Moreover, adding PMN when extracting local features is also helpful to obtain better results.

**Table 3.** Comparison with existing methods on Market1501.

| Method | Market1501 | | | |
|---|---|---|---|---|
| | **Rank-1** | **Rank-5** | **Rank-10** | **mAP** |
| BoW + KISSME [33] | 44.4 | 63.9 | 72.2 | 20.8 |
| WARCA [36] | 45.2 | 68.1 | 76.0 | - |
| SVDNet [34] | 82.3 | 92.3 | 95.2 | 62.1 |
| PDC [38] | 84.4 | 92.7 | 94.9 | 63.4 |
| Triplet Loss [39] | 84.9 | 94.2 | - | 69.1 |
| DML [40] | 87.7 | - | - | 68.8 |
| PAR [13] | 81.0 | 92.0 | 94.7 | 63.4 |
| MFF (Ours) | 96.0 | 98.7 | 99.3 | 87.9 |

The above shows Rank-1 to Rank-5 accuracy (%) and mean Average Precision (mAP) (%).

### 4.4. Comparison with CUHK03

Comparison between the proposed method and CUHK03 is detailed in Tables 4 and 5. We conduct experiments on a CUHK03-detected dataset and a CUHK03-labeled dataset, respectively. We only use the single-query method for person re-identification on CUHK03-detected and CUHK03-labeled datasets. In this paper, our model is compared with many methods, such as LOMO+KISSME [6] using a horizontal occurrence model, pedestrian alignment network [41] and HA-CNN [25] using harmonious attention network. In this experiment, we use Rank-1 accuracy and mAP as evaluation indicators. As shown in Table 4, the MFF model achieves Rank-1 accuracy of 67.4% which is 0.6% higher than the best experimental result on CUHK03-detected data. Additionally, the mAP reaches 66.7%, which is 0.7% better than the best experimental result. Comparison results obtained on CUHK03-labeled are as follows: we surpass MGN by 1.6% in Rank-1 accuracy for the single-query setting. The MFF model reaches mAP of 68.8%. Compared with other deep learning methods, our

model is even more discriminative, which is attributed to our global feature extraction and each-part feature extraction. We believe that local feature extraction benefits from PMN, which is because PMN can extract low-to-high level features more comprehensively.

**Table 4.** Comparison with existing methods on CUHK03-detected data.

| Method | CUHK03-Detected | |
|---|---|---|
| | **Rank-1** | **mAP** |
| BoW + KISSME [33] | 6.4 | 6.4 |
| LOMO + KISSME [6] | 12.8 | 11.5 |
| IDE [42] | 21.3 | 19.7 |
| PAN [41] | 36.3 | 34.0 |
| DPFL [43] | 40.7 | 37.0 |
| SVDNet [34] | 41.5 | 37.3 |
| HA-CNN [25] | 41.7 | 38.6 |
| MLFN [44] | 52.8 | 47.8 |
| PCB+RPP [11] | 63.7 | 57.5 |
| MGN [45] | 66.8 | 66.0 |
| MFF (Ours) | 67.4 | 66.7 |

Rank-1 accuracy (%) and mAP (%) are compared.

**Table 5.** Comparison with existing methods on CUHK03-labeled data.

| Method | CUHK03-Labeled | |
|---|---|---|
| | **Rank-1** | **mAP** |
| BoW + KISSME [31] | 7.9 | 6.4 |
| LOMO + KISSME [6] | 14.8 | 11.5 |
| IDE [42] | 22.2 | 19.7 |
| PAN [41] | 36.9 | 34.0 |
| DPFL [43] | 43.0 | 37.0 |
| SVDNet [34] | 40.9 | 37.3 |
| HA-CNN [25] | 44.4 | 38.6 |
| MLFN [44] | 54.7 | 49.2 |
| MGN [45] | 68.0 | 67.4 |
| MFF (Ours) | 69.6 | 68.8 |

Rank-1 accuracy (%) and mAP (%) are compared.

### *4.5. Comparison with DukeMTMC-reID*

We compare the MFF model with a state-of-the-art model on DukeMTMC-reID. Comparative details are shown in Table 6. Methods of extracting features are different in Table 6, for example, LOMO+KISSME [6] extract local features with a horizontal occurrence model, whereas PAN [41] and SVDNet [34] use a deep learning method to extract global features.

**Table 6.** Comparison with existing methods on DukeMTMC-reID.

| Method | DukeMTMC-reID | |
|---|---|---|
| | **Rank-1** | **mAP** |
| BoW + KISSME [33] | 25.1 | 12.2 |
| LOMO + KISSME [6] | 30.8 | 17.0 |
| Verif + Identif [46] | 68.9 | 49.3 |
| ACRN [47] | 72.6 | 52.0 |
| PAN [41] | 71.6 | 51.5 |
| SVDNet [34] | 76.7 | 56.8 |
| DPFL [43] | 79.2 | 60.6 |
| HA-CNN [25] | 80.5 | 63.8 |
| Deep-Person [48] | 80.9 | 64.8 |
| PCB+RPP [11] | 83.3 | 69.2 |
| MFF (Ours) | 86.0 | 76.1 |

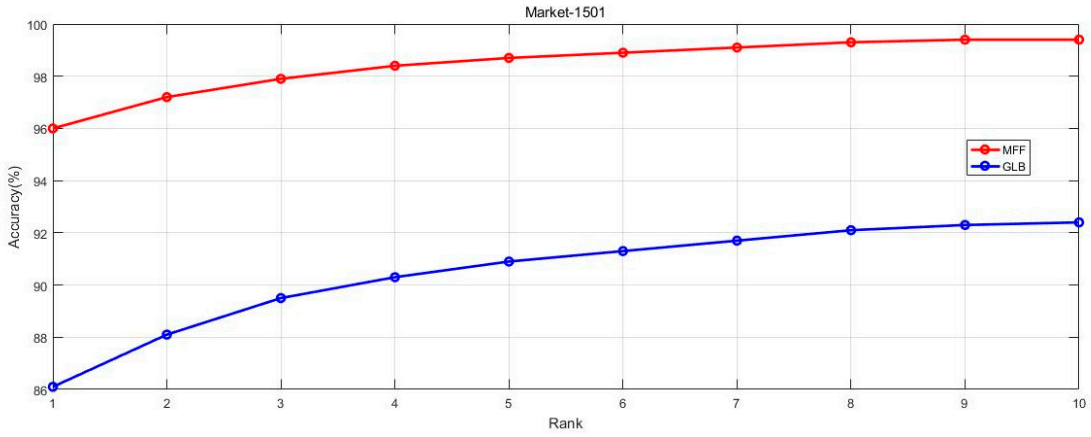
Rank-1 accuracy (%) and mAP (%) are compared above.

We evaluate the MFF model on DukeMTMC-reID with single-query-setting and the significant advantage can be observed in Table 6. Rank-1 accuracy reaches 86.0% which achieves the highest accuracy in comparison methods. We also use mAP as an evaluation indicator. MFF model reaches 76.1% in mAP. Extracting local features and global features enrich the available features when searching for target pedestrians. Adding a classifier in different levels of ResNet50, which is good for extracting part features, can also increase the accuracy of our model. In addition, we visualize the top-10 ranking results on DukeMTMC-reID for some randomly-selected query pedestrian images in Figure 7.

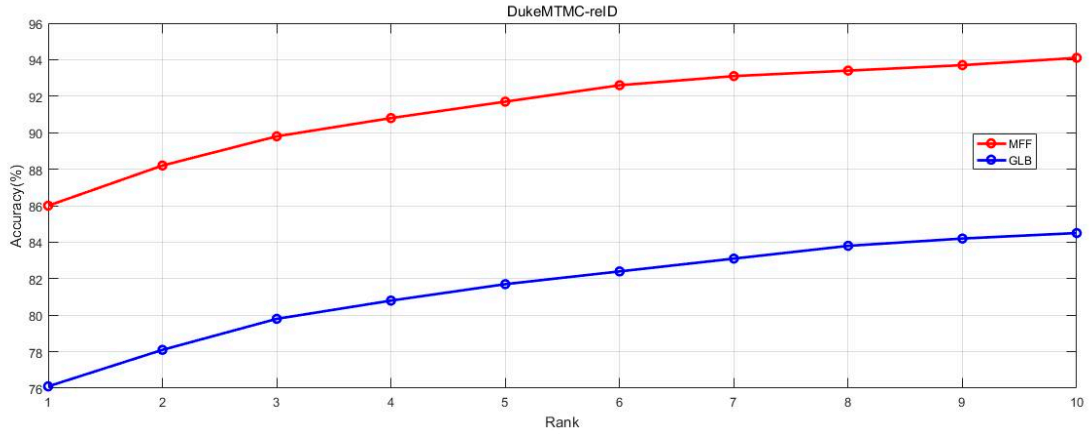

**Figure 7.** One example query image in DukeMTMC-reID dataset and ranking list results from Rank-1 to Rank-10 using MFF model. The blue boundary means true positive and red means false positive.

### *4.6. Effectiveness of PMN*

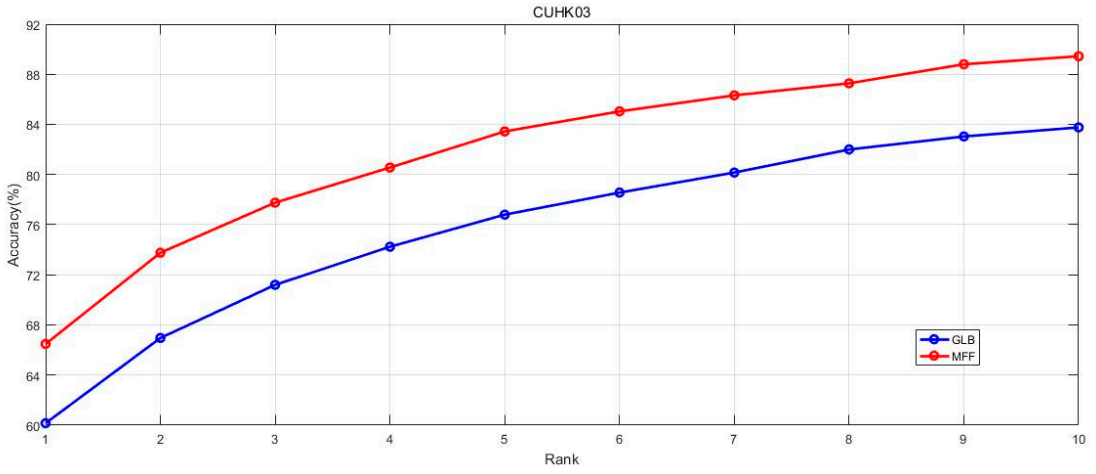
We evaluate the MFF model compared to three classic datasets: Market1501, CUHK03 and DukeMTMC-reID. PMN is proposed to extract local features from the low-to-high level layers. In order to further explore the influence of the PMN model, we conduct two experiments on each dataset. Firstly, we remove the structure of the PMN model. We fuse local features and global features extracted from entire backbone network. GLB is the structure without the PMN model, as in Figure 8. Experiments on GLB can clearly test the performance of our model without adding the PMN structure. Then we train the MFF model on three datasets and report their performance in Figure 8. Difference between MFF and GLB is that MFF fuses low-to-high level local features.




(**a**)

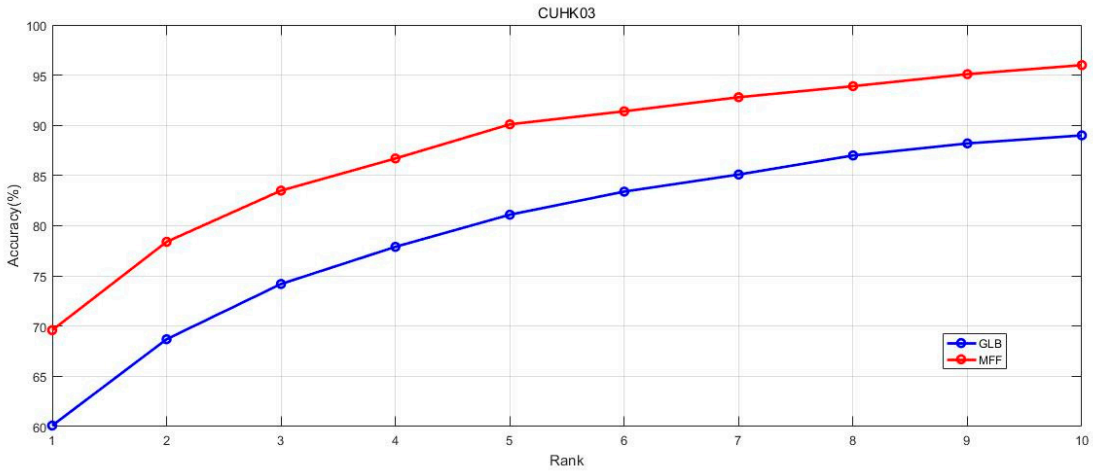



(**b**)




(**c**)

**Figure 8.** *Cont.*




(**d**)

**Figure 8.** Evaluation of GLB and MFF are shown above. Rank-1 to Rank-10 accuracy is compared on three datasets. From (**a**–**d**) is the evaluation performance on Market-1501, DukeMTMC-reID, CUHK03-detected and CUHK03-labeled.

We exhaustively train MFF and GLB on three datasets separately and use Rank-1 accuracy to Rank-10 accuracy as the evaluation standard. In Figure 8, a comparison of experimental results of two models not only shows the effect of model enhancement after fusing low-to-high level local features, but also shows that the improvement effect of PMN on each dataset is different. PMN structure has the most significant effect on CUHK03 especially on CUHK03-labeled data. But the effect on Market-1501 is less significant. Figure 8 shows that rank accuracy of MFF is higher than GLB on three datasets, which proves that low-to-high local features extracted by PMN structure have a positive impact on person re-identification.

### *4.7. Influence of the Number of Parts*

In this paper, we use the method of dividing a pedestrian image into several parts to extract local features. The visualization of the delicate parts is shown in Figure 9. Intuitively, the granularity of the part feature affects the results. When the number of parts is one, the learned feature is a global feature. As the number of divided parts increases, the retrieval accuracy increases. However, accuracy does not always increase with the number parts, as shown in Figure 10. Rank-1 accuracy of three datasets shows that when the number of parts increases to eight, the performance drops dramatically. The over-increased parts actually compromise the extraction of local features. Therefore, we use six parts in our experiments.

**Figure 9.** Visualization of the parts under six values.

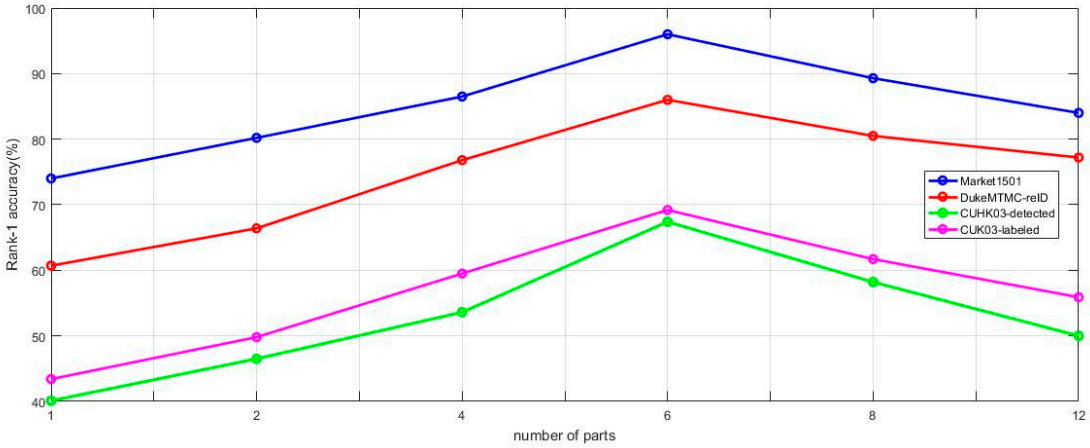


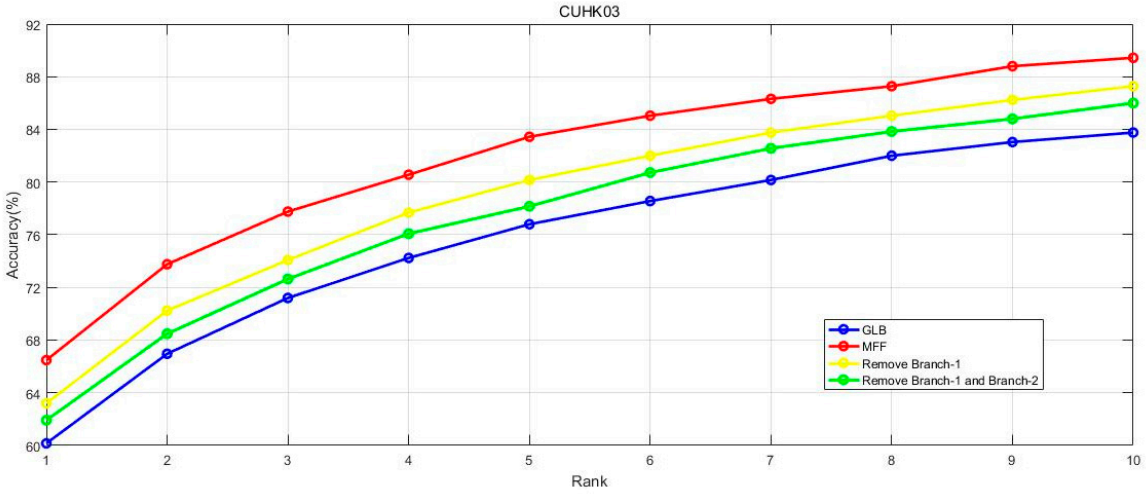

**Figure 10.** Visualization of the parts under six values.

Discussion: we divide the pedestrian image into six parts to get the best results. We consider different proportions and attributes of body parts. We divide the image into six parts according to the position of the elbow joint, crotch, knee joint, etc., as shown in Figure 2. Due to the limitation of joints, the grate range of human motion is limited to these six parts. The image is divided into six parts to ensure that the local features of each part have a high degree of recognition when a pedestrian is engaged in a wide range of activities. In addition, we also consider the effect of attributes on the results. The relevant attributes in pedestrian images include clothing categories (dresses, shorts etc.), clothing color, hat, hair, etc. The recognition of the attribute features of each part is also strengthened after dividing the image into six parts.

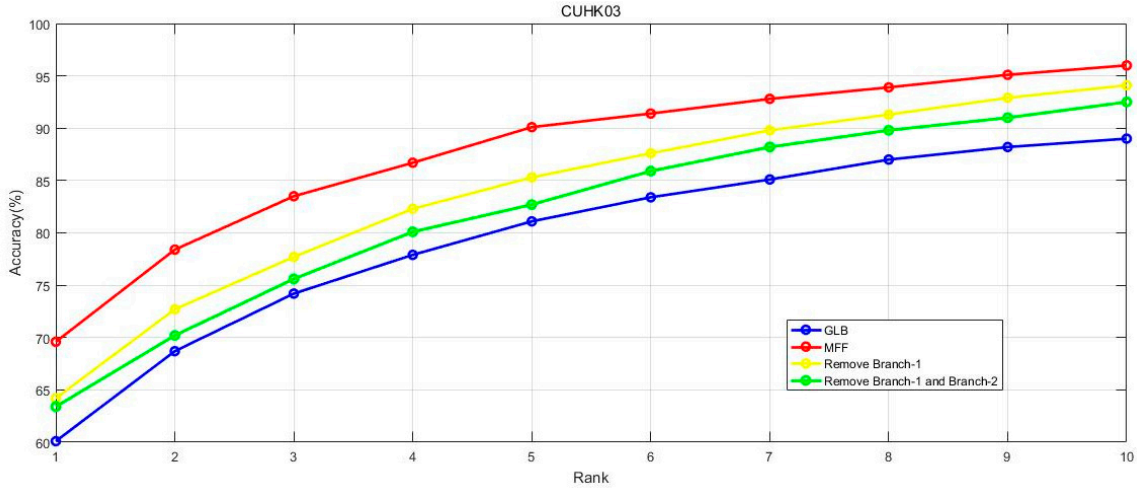
### *4.8. Influence of the PMN Branches*

Low-to-high level local features are extracted by Branch-1 to Branch-3 as in Figure 3. To verify the effectiveness of different branches in PMN, we remove the branches of PMN in different ways and the experimental results are compared in Figure 11. The way of removing branches is as follows. (1) Only Branch-1 is removed. (2) Branch-1 and Branch 2 are both removed. (3) Structure of PMN (Branch-1 to Branch-2) is removed (GLB). (4) No branches are removed (MFF). In Figure 11, we can observe that MFF model achieves the highest rank precision. Removing Branch-1 means not extracting low-level local features which reduces the rank accuracy. In the same way, the more branches in PMN are removed, the lower rank accuracy of the model. This experiment proves that sampling local features from different depths is effective for MFF.

We can try to use PMN networks with different network structures to extract features in the future. In addition, the PMN branches can be used for face recognition to extract facial features from different network depths and learn higher discriminative features. PMN has a wide range of applications and can also be used in other image recognition networks.




(**a**)




(**b**)

**Figure 11.** Impact of low-to-high level local features. Rank-1 to Rank-10 accuracy is compared on datasets CUHK03-detected (**a**) and CUHK03-labeled (**b**).

## 5. Conclusions

This paper mainly verified the important role of our model in solving person re-identification problems. A deep learning network called Multi-level Feature Fusion (MFF) is proposed to extract local features and global features. The proposed Part-based Multi-level Net (PMN) structure not only extracts local features more comprehensively from low to high levels, but also can be flexibly applied into different deep learning models. PMN greatly improves the performance of Multi-level Feature Fusion (MFF) by extracting different levels of local features. A more comprehensive feature fusion effectively improves the accuracy of searching for the target person in person re-identification and outperforms the current state-of-the-art methods with considerable margins.

**Author Contributions:** Conceptualization, S.W.; Methodology, S.W.; Resources, L.G.; Validation, S.W. and L.G.; Writing – original draft, S.W.; Writing – review and editing, S.W. and L.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (61672329).

**Conflicts of Interest:** The authors declare that they have no conflict of interest.

## References


1. 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 Workshops, Salt Lake City, UT, USA, 28 June 2018; pp. 6036–6046.
2. Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 6 September 2018; pp. 365–381.
3. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 24 July 2017; pp. 7291–7299.
4. Gong, K.; Liang, X.; Zhang, D.; Shen, X.; Lin, L. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 16 July 2017; pp. 932–940.
5. Wei, L.; Zhang, S.; Yao, H.W.; Gao, W.; Tian, Q. GLAD: Global-local-alignment descriptor for pedestrian retrieval. In Proceedings of the ACM International Conference on Multimedia, Mountain View, CA, USA, 13 October 2017; pp. 420–428.
6. Liao, S.; Hu, Y.; Zhu, X.; 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 Workshops, Boston, MA, USA, 20 June 2015; pp. 2197–2206.
7. Xiao, Q.; Luo, H.; Zhang, C. Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 7 July 2017.
8. Li, W.; Zhu, X.; Gong, S. Person re-identification by deep joint learning of multi-loss classification. In Proceedings of the International Joint Conferences on Artificial Intelligence, Melbourne, Australia, 1 May 2017; pp. 2194–2200.
9. Yao, H.; Zhang, S.; Zhang, Y.; Li, J.; Tian, Q. Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 2860–2871. [CrossRef] [PubMed]
10. 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 Workshops, Las Vegas, NV, USA, 14 June 2016; pp. 1335–1344.
11. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond Part Models: Person Retrieval with Refined Part Pooling. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 6 September 2018; pp. 480–496.
12. Li, D.; Chen, X.; Zhang, Z. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 25 July 2017; pp. 384–393.
13. Zhao, L.; Li, X.; Wang, J.; Zhuang, Y. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 23 October 2017; pp. 3219–3228.
14. Kalayeh, M.M.; Basaran, E.; Gökmen, M.; Kamasak, M.E.; Shah, M. Human Semantic Parsing for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 20 June 2018; pp. 1062–1071.
15. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-Aligned Bilinear Representations for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 9 June 2018; pp. 402–419.
16. Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 18 June 2014; pp. 152–159.
17. 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, Amsterdam, The Netherlands, 24 October 2016; pp. 791–808.

18. Xiao, T.; Li, H.; Ouyang, W.; Wang, X. Learning deep feature representations with domain guided dropout for person reidentification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, Las Vegas, NV, USA, 26 June 2016; pp. 1249–1258.
19. Yang, Y.; Liu, X.; Ye, Q.; Tao, D. Ensemble Learning-Based Person Re-identification with Multiple Feature Representations. *Complexity* **2018**, *2018*. [CrossRef]
20. Wang, H.; Zhu, X.; Gong, S.; Xiang, T. Person re-identification in identity regression space. *Int. J. Comput. Vis.* **2018**, *126*, 1288–1310. [CrossRef] [PubMed]
21. Paolanti, M.; Romeo, L.; Liciotti, D.; Pietrini, R.; Cenci, A.; Frontoni, E.; Zingaretti, P. Person Re-Identification with RGB-D Camera in Top-View configuration through Multiple Nearest Neighbor Classifiers and Neighborhood Component Features Selection. *Sensors* **2018**, *18*, 3471. [CrossRef] [PubMed]
22. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 18 October 2017; pp. 350–359.
23. Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; Xu, Y. Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 20 November 2017.
24. Liu, H.; Feng, J.; Qi, M.; Jiang, J.; Yan, S. End-to-End Comparative Attention Networks for Person Re-Identification. *IEEE Trans. Image Process.* **2017**, 3492–3506. [CrossRef] [PubMed]
25. Lei, W.; Zhu, X.; Gong, S. Harmonious Attention Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 6 June 2018; pp. 2285–2294.
26. Chen, D.; Yuan, Z.; Hua, G.; Zheng, N.; Wang, J. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 1 June 2015; pp. 1565–1573.
27. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-Margin Softmax Loss for Convolutional Neural Networks. *arXiv* **2016**, arXiv:1612.02295.
28. 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.
29. 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 and Pattern Recognition Workshops, Boston, MA, USA, 10 June 2015; pp. 3908–3916.
30. Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; Sun, J. Alignedreid: Surpassing human-level performance in person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 8 November 2017.
31. 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 Workshops, Honolulu, HI, USA, 6 July 2017; pp. 403–412.
32. Hoffer, E.; Ailon, N. Deep Metric Learning Using Triplet Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 20 December 2015.
33. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person reidentification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 9 December 2015; pp. 1116–1124.
34. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. SVDNet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 16 October 2017; pp. 3800–3808.
35. Engel, C.; Baumgartner, P.; Holzmann, M.; Nutzel, J.F. Person re-identification by support vector ranking. In Proceedings of the British Machine Vision Conference, Aberystwyth, UK, 1 August 2010.
36. Jose, C.; Fleuret, F. Scalable metric learning via weighted approximate rank component analysis. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 1 October 2016; pp. 875–890.
37. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Miami, FL, USA, 7 June 2009.

38. 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, 25 October 2017; pp. 3960–3969.
39. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person reidentification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 17 March 2017.
40. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 1 June 2018; pp. 4320–4328.
41. Zheng, Z.; Zheng, L.; Yang, Y. Person Alignment Network for Large-scale Person Re-identification. *IEEE Trans. Image Process.* **2018**. [CrossRef]
42. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-identification: Past, Present and Feature. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 10 October 2016.
43. Chen, Y.; Zhu, X.; Gong, S. Person re-identification by deep learning multi-scale representations. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 15 September 2017; pp. 2590–2600.
44. Chang, X.; Hospedales, T.M.; Xiang, T. Multi-Level Factorisation Net for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 24 March 2018; pp. 2109–2118.
45. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning Disriminative Features with Multiple Granularities for Person Re-Identification. *arXiv* **2018**, arXiv:1804.01438.
46. Zheng, Z.; Zheng, L.Y.; Yang, Y. A Discriminatively Learned Cnn Embedding for Person Re-identification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 13. [CrossRef]
47. Schumann, A.; Stiefelhagen, R. Person Re-identification by Deep Learning Attribute-Complementary Information. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 8 October 2017; pp. 20–28.
48. Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; Xu, Y. Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. *arXiv* **2017**, arXiv:1711.10658. [CrossRef]