algorithms

MDPI

Article

Multi-Level Fusion Model for Person Re-Identification by Attribute Awareness

Shengyu Pei [1] and Xiaoping Fan [1,2,*]

[1] School of Automation, Central South University, Changsha 410075, China; shengyupei@csu.edu.cn
[2] School of Information Technology and Management, Hunan University of Finance and Economics, Changsha 410205, China
* Correspondence: xpfan@csu.edu.cn

Abstract: Existing person re-recognition (Re-ID) methods usually suffer from poor generalization capability and over-fitting problems caused by insufficient training samples. We find that high-level attributes, semantic information, and part-based local information alignment are useful for person Re-ID networks. In this study, we propose a person re-recognition network with part-based attribute-enhanced features. The model includes a multi-task learning module, local information alignment module, and global information learning module. The ResNet based on non-local and instance batch normalization (IBN) learns more discriminative feature representations. The multi-task module, local module, and global module are used in parallel for feature extraction. To better prevent over-fitting, the local information alignment module transforms pedestrian attitude alignment into local information alignment to assist in attribute recognition. Extensive experiments are carried out on the Market-1501 and DukeMTMC-reID datasets, whose results demonstrate that the effectiveness of the method is superior to most current algorithms.

Keywords: person re-identification; attribute awareness; non-local; multi-task

Citation: Pei, S.; Fan, X. Multi-Level Fusion Model for Person Re-Identification by Attribute Awareness. *Algorithms* **2022**, *15*, 120. https://doi.org/10.3390/a15040120

Academic Editors: Frank Werner and Xingjuan Cai

Received: 9 March 2022
Accepted: 29 March 2022
Published: 30 March 2022

1. Introduction

Given a pedestrian image, the purpose of person re-identification is to retrieve images of the pedestrian from cross-camera devices. Person re-identification is an image-retrieval technology designed to compensate for the visual limitations of fixed cameras. It can be combined with pedestrian-detection and tracking technologies for use in fields such as intelligent pedestrian detection and intelligent security [1–4].

Pedestrians have both rigid and flexible characteristics, due to the differences between camera devices, and their appearance is easily affected by factors such as clothing, posture, weather, and occlusion, making person re-identification one of the most challenging research topics in the field of computer vision.

The main idea of traditional image-based person re-recognition is to compare the similarity of two identities. The similarity of different identities is small, and the similarity of the same identity is large. Though the supervised person re-identification problem has labeled information, for most person Re-ID datasets, their boundary boxes are detected by outdated detectors. These limitations make it necessary to improve discriminative performances in person Re-ID. Pedestrian attributes have many commonalities between identities, and there are many differences in pedestrian images with the same identity. As a result, the features extracted by the network often cannot accurately measure the similarity. Therefore, simply relying on the labeled information to determine the pedestrian distance can easily lead to the deviation of the network's attention features. For example, assume three pedestrians have identities (x_1, y_1), (x_2, y_2), and (x_3, y_3), where $y_1 \neq y_2 \neq y_3$, while x_1 and x_2 may be very similar, the similarities $S(x_1, x_2)$, $S(x_1, x_3)$, and $S(x_2, x_3)$ will be far apart. As a result, the network pays attention to other regional feature information. When attributes are introduced, (x_1, y_1) can be expressed as (x_1, y_1, A_1). When calculating

the similarity, both the label and the attribute information will be considered to guide the network to pay attention to the corresponding features. Thus, we can conclude that $S(x_1, x_2) > \max(S(x_1, x_3), S(x_2, x_3))$. Thus, we can conclude that attributes help guide the network to learn the relationship between attributes and features and obtain features with more semantic information. Attributes can also speed up training. Adding attributes can filter out some images that are incompatible with query attributes.

Of course, more attributes do not imply better performance. Two classic image-based person re-recognition datasets, Market-1501 and DukeMTMC-reID, have been marked with attribute information. Market-1501 has 30 attributes for each identity, and DukeMTMC-reID has 23. These datasets have many cases of pedestrians with different identities, but whose attributes are very similar. That is, suppose two pedestrians have identities (x_1, y_1, A_1) and (x_2, y_2, A_2), where $A_1 = A_2$ and $y_1 \neq y_2$. It is not possible to directly use attribute labels alone to train the network or to choose discriminative attributes. Therefore, we use identity and attribute labels to jointly train the network. Although current single-modal algorithms have achieved good performance on some standard datasets, there is a long way to go. As we all know, RGB images are sensitive to certain weather conditions such as rain, snow, and fog. However, intelligent monitoring must meet actual needs. The intuitive idea is to mine useful information from other modalities (such as thermal and depth sensors) and fuse it with RGB sensors. Therefore, there has been much work on deep fusion of multi-modal data to improve performance [5]. Scholars have also proposed to introduce attributes to person re-identification tasks. Lin et al. [6] proposed an attribute person recognition (APR) network, a multi-task network that simultaneously learns pedestrian ID embedding and predicts pedestrian attributes. They manually marked attribute labels of two person re-identification datasets and systematically studied the correlation between pedestrian ID and attribute recognition. Yin et al. [7] proposed an identity-recognition network (IRN) and attribute-recognition network (ARN), and used IRN to extract partial information on pedestrians and ARN to calculate their attribute similarity. Due to the role of attributes in detection and recognition, some scholars have introduced them to video-based person re-recognition.

However, at present, person re-identification with attributes still encounters many problems: unbalanced data distribution and low-quality pedestrian images [8,9]. We propose a hybrid network framework and make the following contributions.

1. Our model includes a multi-task learning module, local information alignment module, and global information learning module. The local information alignment module transforms pedestrian attitude alignment into local information alignment to inference pedestrian attributes.
2. We design an improved network based on non-local and instance batch normalization (IBN) to learn more discriminative feature representations.
3. The proposed method outperforms the latest person re-identification methods.

The remainder of this paper is arranged as follows. Section 2 introduces related work of image-based person re-identification. Section 3 introduces our proposed method. Section 4 discusses simulation experiments using the proposed method on two image-based person re-recognition datasets. Section 5 summarizes our proposed method.

2. Related Work

Person re-identification solves the problem of matching pedestrian images between unrelated cameras. It faces challenges caused by different perspectives, postures, occlusion, and other issues. To solve these problems, we need to increase the inter-class distance and reduce the intra-class distance. Traditional methods depend on metric learning [10–12] and deep learning [13,14].

Learning each attribute independently is an intuitive idea, but it makes person attribute recognition redundant and inefficient. Therefore, researchers tend to evaluate all attributes in a network model, and evaluate each attribute as its own task. Due to the high efficiency of multi-task learning, researchers have been paying increased attention to it [15–18].

Lin et al. [6,7] proposed a network framework combining pedestrian ID labels and attributes, breaking through the traditional limitation of learning only using pedestrian ID labels. They construct a multi-task network that simultaneously learns pedestrian ID labels and predicts attribute labels by introducing pedestrian attribute labels. Person re-identification is challenging because pedestrians are both flexible and rigid, and camera devices are different and greatly affected by the environment. These multi-task networks have high accuracy in attribute recognition, but do not perform person re-identification reliably. Unlike these networks, to improve the accuracy of person re-identification, this paper uses pedestrian attribute labels as an aid, and discusses their function in pedestrian recognition.

Lin et al. [6] and Yin et al. [7] build a multi-task network by introducing pedestrian attribute labels, enabling the network to learn identity and attributes. Lin et al. labeled pedestrian attributes and combined attribute learning and global learning of pedestrian re-identification methods. However, there has been no quantitative analysis of attributes, and the problem of hard samples is not considered. Yin et al. added hard-sample learning to force the network to extract more advanced semantic features. However, the network lacks local learning capability and has limited improvement effects. In this paper, we design a multi-level fusion model with joint local and global learning and introduce various tricks to improve the learning ability, improving the discriminative ability of the network as a whole.

Figure 1 shows an example. Figure 1a shows the role of attributes in pedestrian re-recognition. Through ID tag learning and attribute learning, the network can learn semantic information of pedestrian attributes. However, pedestrians are flexible and rigid, which makes it difficult to re-identify pedestrians. For example, Figure 1b,c. Figure 1b show two different identities. Each identity has two pictures, but they have different points. For example, in the first identity, pedestrians sometimes carry a bag and sometimes do not. In identity 2, sometimes pedestrians ride bikes and sometimes they walk. Figure 1c shows three different identities. Their jackets are all yellow with similar characteristics. If you only use ID tags for learning, it is hard to tell these identities apart. The first identity and the second identity are both male, and the third identity is female. Attribute learning can separate them. The second identity is wearing a backpack and riding a bike. Attribute learning can separate these, too.

Zhu et al. [18] used attributes to assist a person re-identification network, fusing the low-level feature distance and attribute-based distance as the final distance to distinguish whether a given image has the same identity.

Attributes have been introduced to video-based person re-recognition because of their role in detection and recognition. Zhao et al. [19] proposed an attribute-driven method for feature decomposition and frame weighting. The sub-features are re-weighted through the confidence of attribute recognition and integrated into the time dimension as the final representation. Through this strategy, the area with the largest amount of information in each frame is enhanced, which contributes to a more differentiated sequence representation. Song et al. [20] proposed the partial attribute-driven network (PADNet). Methods such as this are based on global-level feature representation. Pedestrians are automatically divided into multiple body parts. A four-branch multi-label network is used to explore the spatiotemporal cues of the video.

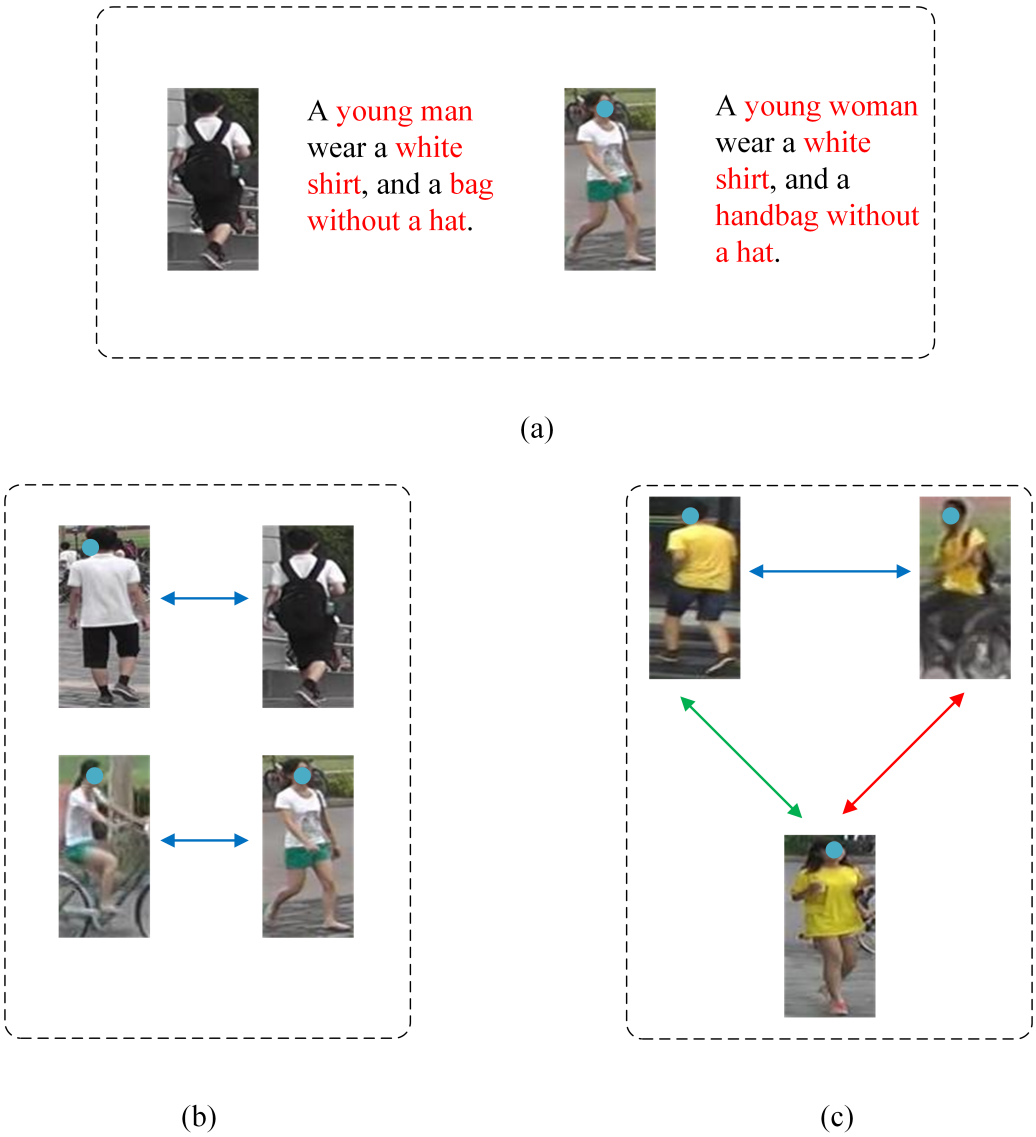


Figure 1. Pedestrian flexibility and rigidity bring challenges to pedestrian re-recognition. (**a**) Learning the semantic information of pedestrian attributes. The red text represents some features of the pedestrian. (**b**) The different states of two different pedestrians. (**c**) The distance between different pedestrians. The double arrows show their distance. Different colors indicate the distance between different pedestrians, and these distances need to be pulled apart.

The bulk of person re-identification is based on static images. Although the ideas around solving the problem are different, the aim is to retrieve the most similar image. In the training phase, the distances of the same class should be as close as possible, and the distances of different classes should be separated as much as possible. In the testing phase, we compare all pedestrian images in the gallery, and select the one with the closest distance. Translated into the problem of retrieving the most similar image, the construction of features is particularly important. If we treat this problem according to the traditional artificial perspective, we will judge the identity of pedestrians by criteria such as the clothing, age, and body.

3. Proposed Method

3.1. Network Structure

We describe the network structure in detail.

Figure 2 shows the proposed network framework, which has two parts: attribute recognition and identity recognition, corresponding to A and B, respectively. A and B have parts in common. Features of attribute identification and the global branch are derived from "Feature Aggregation 1". The features of the local branch are derived from "Feature Aggregation 2". Unlike the global branch, the local branch needs to produce features of six different parts.

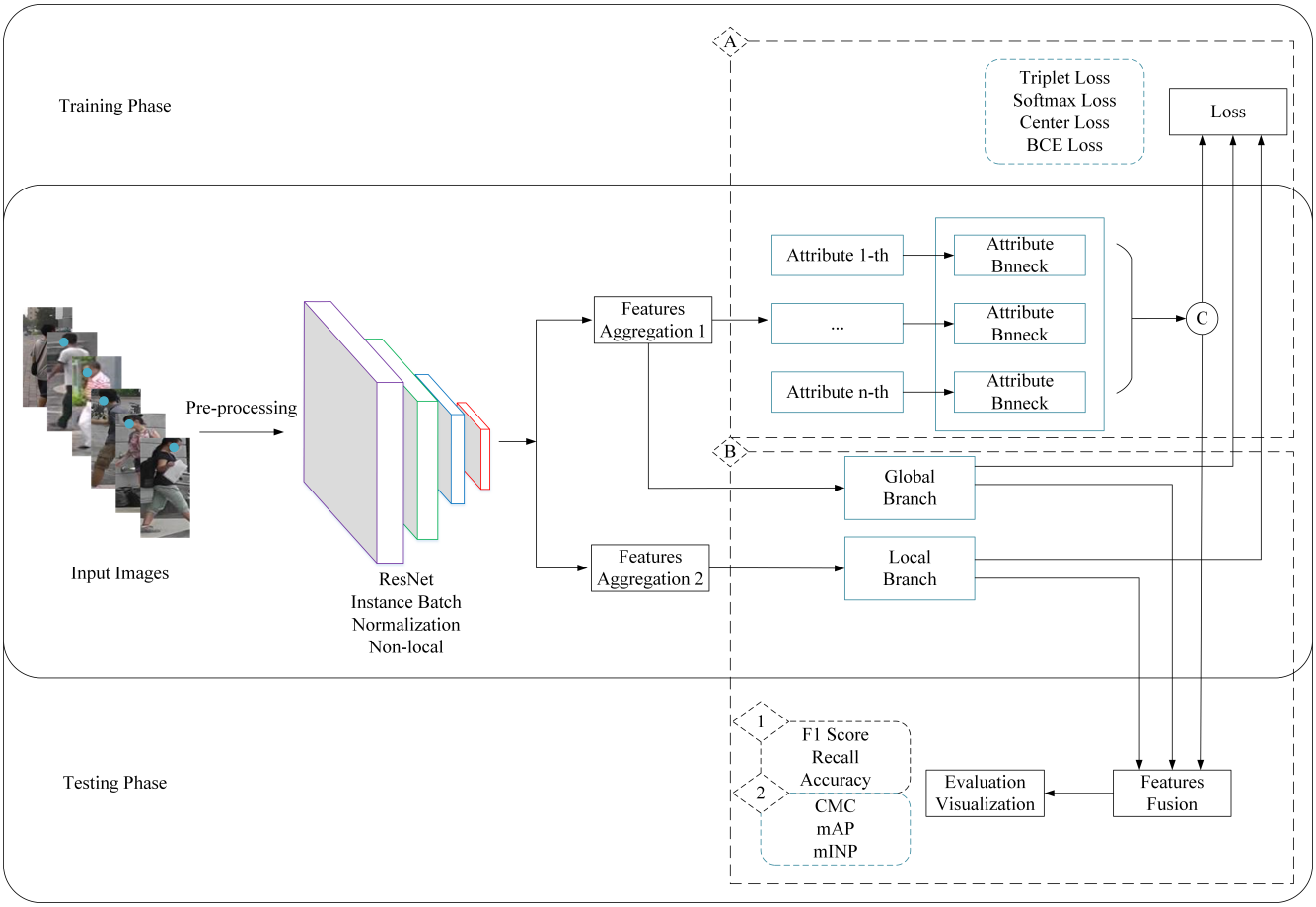


Figure 2. Network structure. It contains two parts: the training and testing phase. The middle part of the figure represents the common part of training and testing. It contains the pedestrian identity and attributes recognition. The training part needs to calculate the loss function, and the model weights are updated using backpropagation. In the test phase, the results of each indicator are output through feature fusion. Backbone uses different colors to distinguish the four different network layers of ResNet.

We separate the identity network from the entire network framework to understand the structure more clearly, as shown in Figure 3. Figure 3 is a multi-level feature-fusion network focusing only on identity recognition. It is a person Re-ID network framework without attribute recognition. Figure 3 contains the global feature-extraction module and local feature-extraction module. The goal of the global module is to extract global information on pedestrians. The purpose of the local module is to align the pedestrian within the boxes. After the local average pooling layer, six different part-level features' expressions are generated.

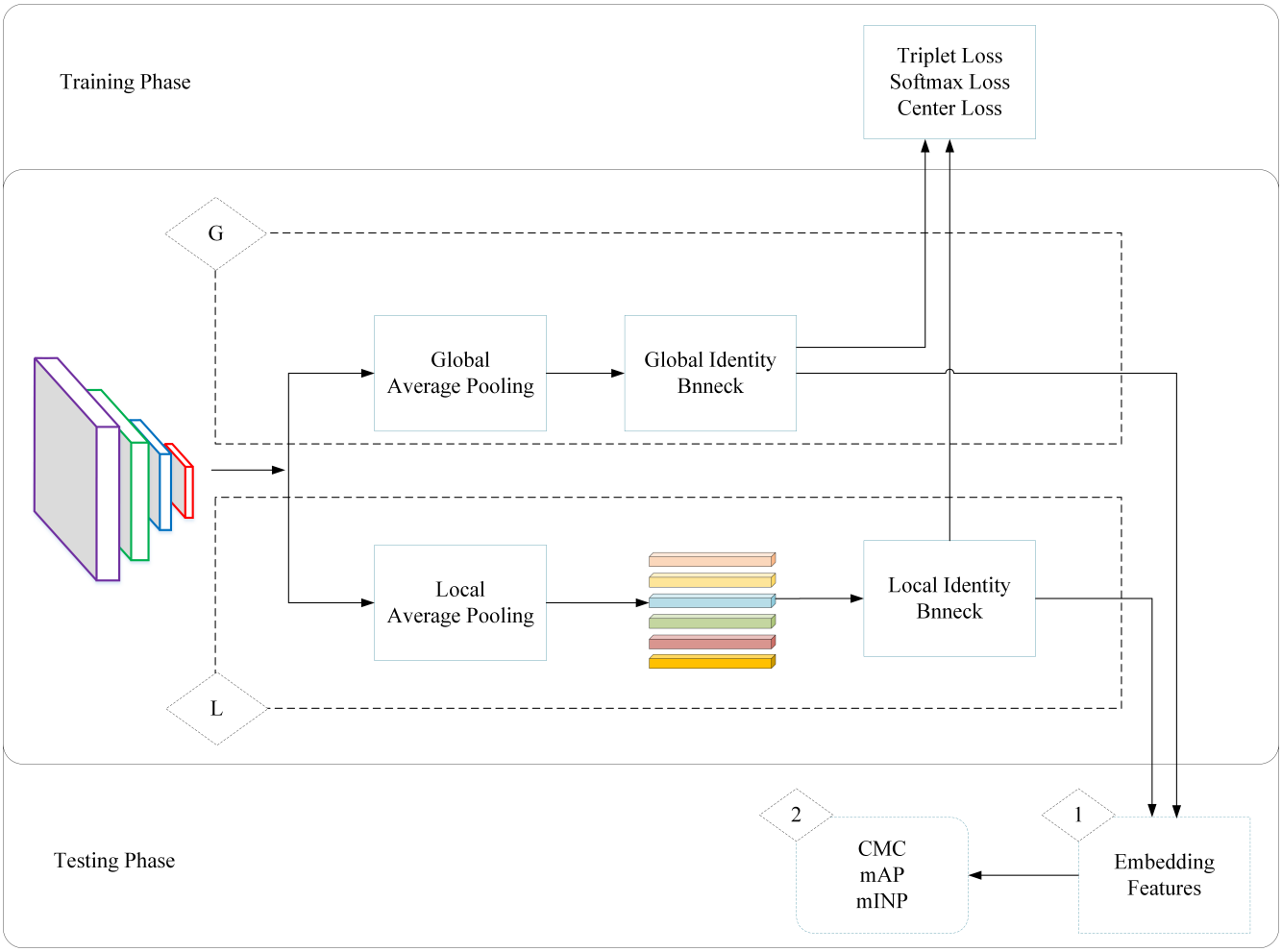


Figure 3. Global and local branch network, which is a multi-level feature-fusion network. It is a pedestrian re-recognition network framework without attribute recognition. After local average pooling, six different part features are generated, represented by different colors in this figure.

In the training phase, the attribute-identification branch, the result of the global branch and the local branch calculate the loss value through the loss function, and complete backpropagation. The "C" in module A of Figure 2 represents a concat function to combine the classification results of all attributes. BCE loss is used to calculate the difference between the predicted and real attributes and to learn the relationship between pedestrian attributes and aggregated features through backpropagation.

In the testing phase, the results of the global and local branches are fused, and the required results are output by an evaluation function. Points "1" and "2" of Figure 3 represent the output results of person re-identification with attributes in the testing phase. The metrics in "1" of Figure 3 are commonly used for person re-identification with attributes. It contains F1 score, recall and accuracy.

To provide a clearer backbone to what is proposed in this paper, Figure 3 shows a pedestrian re-recognition network framework without attribute recognition. The framework describes the workflow of the global and local branches in more detail. "G" in the figure represents the global branch and "L" represents the local branch.

In the training phase, we preprocess the input images, which plays the role of data enhancement. The preprocessing operation includes five data-enhancement modules: resize, random horizontal flip, pad, random crop, and random erasing. These can help prevent the network from falling into local extrema during the training process, which leads to overfitting.

Preprocessing also helps realize the diversification of input images and helps the to better train the network. Then, preprocessed data are fed into the ResNet backbone network used in this paper. Two modules, instance batch normalization and the non-local network (Section 3.1), endow ResNet with a stronger feature-extraction performance.

We introduced generalized mean pooling [21], whose function is similar to that of adaptive average pooling, to aggregate the feature embedding after the backbone. Its mathematical formula is as follows:

$$f_i = \left(\frac{1}{|X_i|} \sum_{x \in X_i} x^{p_i} \right)^{\frac{1}{p_i}} \tag{1}$$

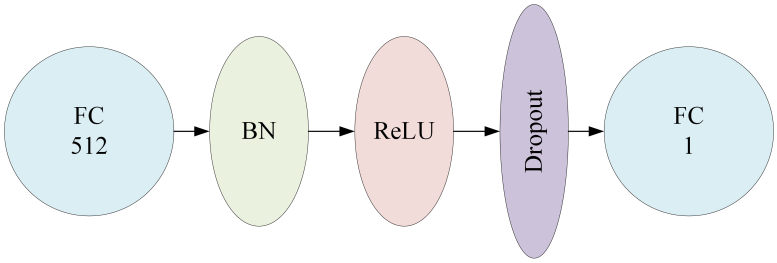
where f_i represents the i-th dimension of feature aggregation, and $f = [f_1 ... f_i ... f_K]$ represents the aggregated features. In particular, when $p_i = \infty$, then generalized mean pooling evolves into max pooling, and when $p_i = 1$, it evolves into average pooling. In this paper, we set $p_i = 3$.

The aggregated features are sent to modules A and B in Figure 2, respectively. Module A learns the relationship between pedestrian attributes and features, and can learn the correlation between attributes and features. Module B learns the relationship between pedestrian identities and features. It contains a global and local branch. Module B can learn the global and part-level features of pedestrian images. Modules A and B combine to focus the network on pedestrian attributes as well as identities, and can prevent network overfitting.

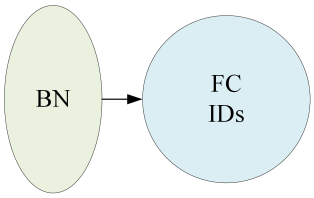
To obtain the relationship between each attribute and the aggregation features, a batch-normalization (BN-1) module is set in module A for each attribute. The BN-1 module is shown in Figure 4a, followed by a two-classifier, to determine whether the current pedestrian feature contains this attribute.

In module B in Figure 2, to obtain the relationship between each identity and the aggregated features, a batch-normalization (BN-2) module (Figure 4b) is designed. This has a similar function to BN-1 in Figure 4a, but only a simple one-dimensional batch normalization operation is used. Finally, triplet, softmax, and center loss are used to calculate the difference between the predicted and real identities. Using backpropagation to learn the relationship between pedestrian identity and aggregation features, the network brings pedestrians of the same identity as close as possible, and keeps pedestrians of different identities farther apart.

The testing phase differs from the training phase. The network does not perform data enhancement on the input images, but to adapt to the network, it performs a resizing operation. After the aggregated features are obtained, the attributes of the current input images (query and gallery) can be judged through the attribute classifiers. The network can also output the embedded features of each input image (query) through the BN-2 module in Figure 4b, and find the images with the best rank score from the gallery through the distance-matching method. The "1" and "2" in Figure 3 represent the output results of the double-stream network in the testing phase. The metrics in "2" are commonly used for person re-identification problems. The embedded features in "1" are used for output visualization.



(a) Batch normalization (BN-1)



(b) Batch normalization (BN-2)

Figure 4. Batch normalization modules are used for different tasks. We design two modules: BN-1 for attribute recognition and BN-2 for identity recognition. "IDs" is the number of all different identities.

3.2. Non-Local Residual Network (ResNet) of Instance Batch Normalization (IBN)

We add attention-like non-local [22] and instance batch normalization (IBN) [23] modules to learn more robust features.

The generalized non-local operation can be defined as:

$$z_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{2}$$

where i represents the output location and j represents all possible locations, x is the input image, and $f(x_i, x_j)$ is a function that calculates the affinity of i and j. g is a unary function representing the input image at location j, and $C(x)$ is a normalization function.

We apply non-local operations to the ResNet of instance batch normalization, making some changes to adapt it. Specifically, $C(x) = N$, the number of locations of x, and $f(x_i, x_j)$ uses a dot product to calculate the affinity of i and j,

$$f(x_i, x_j) = \theta(x_i)^T \phi(x_j) \tag{3}$$

In this paper, the non-local block acts on Layer 2, Layer 3 and Layer 4 of ResNet.

3.3. Loss Function

In a person re-identification dataset, we can use $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ to represent the set with identity labels, where x_i and y_i are the input image and label, respectively, of identity i, and n is the total number of input images [24,25]. In the person re-identification dataset with attributes, we can use A to represent the attribute set of all identities, where $A_i = (A_i^1, A_i^2, ..., A_i^m)$ is the attribute subset of identity i, and m is the number of attributes of each identity. Therefore, we can use $E = \{(x_1, A_1), (x_2, A_2), ..., (x_n, A_n)\}$ to represent a set with attribute labels. For the two sets D and E, we use a bidirectional parallel approach to solve the person re-identification problem [26,27]. Therefore, we can define the following three functions.

For the set D with identity labels, we define two functions, both based on the objective function of identity labels. First is a classification function based on identity labels,

$$L_{Id} = -\sum_{i=1}^{n} y_i \log q_i \tag{4}$$

$$F_{Id} = \min_{w_{Id},\theta} \sum_{i=1}^{n} L_{Id}(f_{Id}(w_{Id}, \phi(\theta, x_i)), y_i) \tag{5}$$

where $q_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$ is the confidence probability of the prediction label. $\phi(\theta, x_i)$ is the feature embedding function of the first ith identity. θ is the training parameter of the feature embedding function. $f_{Id}(w_{Id}, \phi)$ is the classification function of the feature that embeds the identity label, and w_{Id} is the training parameter of the classification function. $L_{Id}(f_{Id}, y_i)$ is the image label loss function of identity i. The purpose of the function F_{Id} is to find the appropriate image feature embedding so that the identity obtained by training is as consistent as possible with the real label.

The second is a metric learning function based on identity labels.

$$L_{Tri} = \max([d_p - d_n + \alpha], 0) \tag{6}$$

$$F_{Tri} = \min_{\theta} \sum_{i=1}^{n} L_{Tri}(\phi(\theta, x_i), y_i) \tag{7}$$

where d_p is the Euclidean distance of the positive pairs, d_n is the Euclidean distance of the negative pairs, α is the decision boundary of the Triplet loss. $L_{Tri}(\phi, y_i)$ is the metric loss function of the feature embedding identity. The purpose of F_{Tri} is to find the appropriate image feature embedding so that the identities of the same labels are as close as possible, and those of different labels are separated as much as possible.

We combine classification learning and metric learning to find an image feature embedding to better solve the person re-identification problem. For a set E with attribute labels, we define a function,

$$L_{Att} = -\sum_{m=1}^{M} y_m log(p(m|x)) \tag{8}$$

$$F_{Att} = \min_{w_{Att},\theta} \sum_{i=1}^{n} \sum_{j=1}^{m} L_{Att}(f_{Att_j}(w_{Att_j}, \phi(\theta, x_i)), A_i^j) \tag{9}$$

where y_m is the target class of the m-th attribute, $p(m|x)$ is the confidence probability of the prediction attribute. $f_{Att_j}(w_{Att_j}, \phi)$ is the classification function of the feature that embeds the attribute of identity j, and $L_{Att}(f_{Att_j}, A_i^j)$ is the j-th classification loss function of identity i. We integrate all the attributes of identity i to obtain its attribute set. The purpose of F_{Att} is to find a suitable image feature embedding so that the identity attribute set obtained by training is as consistent as possible with the real attribute set.

In the testing phase, we use the feature embedding function $\phi(\theta, x_i)$ to embed the query set and all gallery set images in the feature space. The identity label of the query image is judged according to the Euclidean distance between each query image and all gallery images, and $f_{Att_j}(w_{Att_j}, \phi(\theta, x_i))$ calculates all the attributes of each query image identity.

To better adapt these classifiers (f_{Id} and f_{Att}), we normalize them after executing the feature embedding function $\phi(\theta, x_i)$. In particular, in the testing phase, we normalize them before calculating the Euclidean distance between each query image and all gallery images.

4. Experiment

We conducted experiments to verify the effectiveness of Algorithm 1. In order to distinguish it from others, we call the method proposed in this paper multi-level model for person re-identification by attribute awareness (MLAReID).

Algorithm 1 MLAReID algorithm.

Input: Initialize learning rate (lr = 0.00035), optimizer ("Adam"), batchsize = 64
Input: Input pedestrain images, pedestrain attributes
Input: Initialize multi-level fusion model (global-local, non-local, IBN)
1: **for** each $i \in [1, epochs]$ **do**
2: Extract feature vectors from input images by the model
3: Predict labels, attributes from input images by the model
4: Update ID loss with Equation (3)
5: Update Triplet loss with Equation (5)
6: Update Attribute loss with Equation (9)
7: **end for**
Output: F1 score, Recall, Accuracy, cmc, mAP, mINP

4.1. Datasets and Settings

1. Market-1501 [28]
   This dataset was collected by six cameras in front of a supermarket at Tsinghua University. It has 1501 identities and 32,668 annotated bounding boxes. Each annotated identity appeared in at least two cameras. The dataset is divided into 751 training identities and 750 testing identities, corresponding to 12,936 and 19,732 images, respectively. Attributes are annotated by pedestrian identity. Each image has 30 attributes. Note that although the upper- and lower-body clothing have seven and eight attributes, respectively, each identity has only one color marked "yes".
2. Duke Multi-Target, Multi-Camera (DukeMTMC-reID) [29,30]
   The dataset from Duke University contains 1812 identities and 34,183 annotated bounding boxes. It is divided into 702 training identities and 1110 testing identities, corresponding to 16,522 and 17,661 images, respectively. Attributes are annotated by pedestrian identity. Each image has 23 attributes.

For the fairness of comparison, each image has a width of 128 pixels and a height of 256 pixels.

4.2. Evaluation Metrics

To measure the performance of the algorithm, we used standard metrics including cumulative matching cure (CMC), mean average precision (mAP), mean inverse negative penalty (mINP), and receiver operating characteristic (ROC) curve.

4.3. Datasets and Settings

The algorithm in this paper uses data-enhancement methods such as random horizontal flip, pad, random crop, and random erasing to preprocess the input images. For the triplet loss function, in the training phase, four identities are fed into the network in each batch, and each identity has eight images, for a total of 32 pedestrian images.

4.4. Comparison with the State-of-the-Art

We compare our method with methods published in recent years. Tables 1 and 2 compare the Rank1, Rank5, Rank10, and mAP evaluation metrics of the Market-1501 and DukeMTMC-reID datasets, respectively, to those of these other methods, where "-" means there is no record.

Table 1. Comparison with the state-of-the-art algorithms on Market-1501.

Method	Rank1	Rank5	Rank10	mAP
MBC [31]	45.56	67	76	26.11
SML [32]	45.16	68.12	76	-
SL [33]	51.9	-	-	26.35
Attri [34]	58.84	-	-	33.04
S-CNN [35]	65.88	-	-	39.55
2Stream [36]	79.51	90.91	94.09	59.84
Cont-aware [37]	80.31	-	-	57.53
Part-align [38]	81.0	92.0	94.7	63.4
SVDNet [39]	82.3	92.3	95.2	62.1
GAN [30]	83.97	-	-	66.07
EBB [40]	81.2	94.6	97.0	-
DSR [41]	82.72	-	-	61.25
AACN [42]	85.90	-	-	66.87
APR [6]	87.04	95.10	96.42	66.89
PN-GAN [43]	89.4	-	-	72.6
CLSA [44]	88.9	-	-	73.1
HAP2S [45]	84.59	-	-	69.43
PABR [46]	90.2	96.1	97.4	76
PCB [47]	92.3	97.2	98.2	77.4
PSE [48]	87.7	94.5	96.8	69
DistributionNet [49]	87.26	94.74	96.73	70.82
DRAL [50]	84.2	94.27	96.59	66.26
AttKGCN [51]	94.4	98	98.7	85.5
Yin [7]	92.8	97.5	98.3	79.5
SCSN (4 stages) [52]	92.4	-	-	88.3
SIAMH [53]	95.4	-	-	88.8
Jin [24]	94.6	-	-	87.5
Zhou [54]	94.8	-	-	86.7
Li [55]	95.5	-	-	88.5
MLAReID	96.1	98.5	99.3	90.3
MLAReID + Reranking	96.5	98.2	98.8	95.4

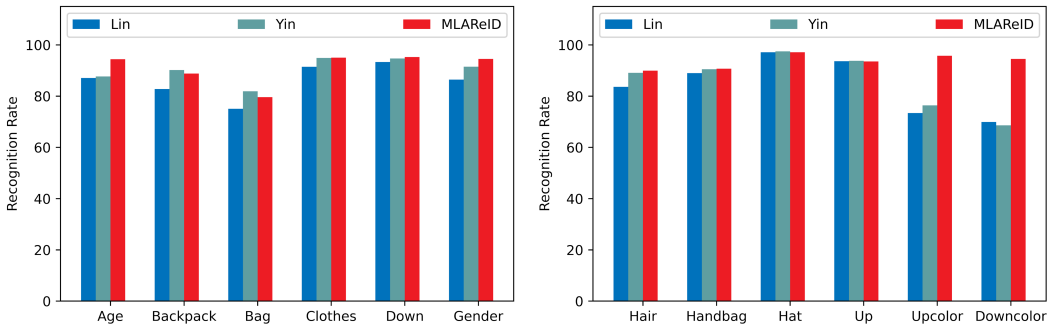
For the Market-1501 dataset, the methods compared are GAN-based, part-based, and combined. From Table 1, we find that our proposed method achieved the best results on Rank1, Rank5, Rank10, and mAP when compared with current state-of-the-art methods.

Table 2. Comparison with the state-of-the-art algorithms on DukeMTMC-reID.

Method	Rank1	Rank5	Rank10	mAP
BoW + kissme [28]	25.13	-	-	12.17
LOMO + XQDA [56]	30.75	-	-	17.04
AttrCombine [34]	53.87	-	-	33.35
GAN [30]	67.68	-	-	47.13
SVDNet [39]	76.7	-	-	56.8
APR [6]	73.92	-	-	55.56
PSE [48]	79.8	89.7	92.2	62
DistributionNet [49]	74.73	85.05	88.82	55.98
AttKGCN [51]	87.8	94.4	95.7	77.4
Yin [7]	82.7	91	93.5	66.4
SCSN (4 stages) [52]	91.0	-	-	79.0
SIAMH [53]	90.1	-	-	79.4
Jin [24]	88.6	-	-	78.4
Zhou [54]	88.7	-	-	76.6
Li [55]	90.2	-	-	79.7
MLAReID	91.4	95.5	96.7	81.4
MLAReID + Rerankingg	92.7	96.1	97.2	90.6

For the DukeMTMC-reID dataset, compared with other methods on the evaluation metrics Rank1, Rank5, Rank10, and mAP, the method proposed in this paper achieves the best performance.

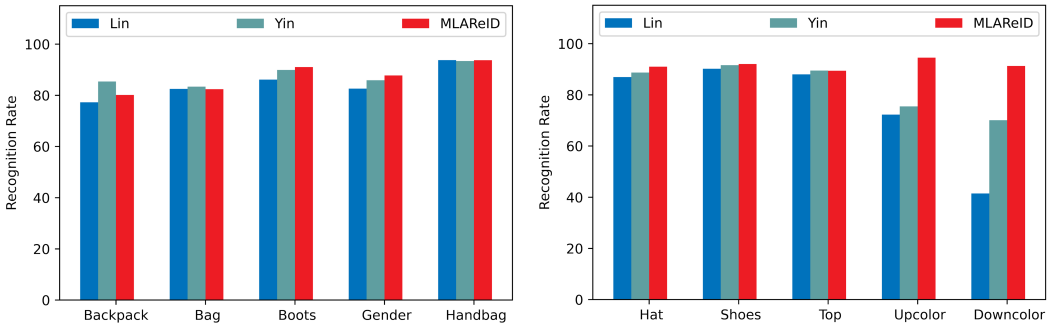
For the Market-1501 dataset, this paper lists the accuracy of 12 attributes for comparison. "age" has four attributes: young, teenager, adult, and old. "upcolor" has eight attributes: upblack, upwhite, up-red, uppurple, upyellow, upgray, upblue, and upgreen. "downcolor" has nine attributes: downblack, downwhite, downpink, downpurple, downyellow, downgray, downblue, downgreen, and downbrown. "Upcolor" and "Downcolor" in Figure 5b represent the average accuracy of these two sets of related attributes. It can be found from Figure 5 that recognition rate of "Backpack", "Bag", "Hat", and "Up" are not as good as the method proposed by Yin [7], but the recognition rates of the other attributes are significantly improved.



(a) Attribute recognition rate a (b) Attribute recognition rate b

Figure 5. Attribute recognition accuracy on Market-1501 dataset: (**a**) age, backpack, bag, clothes, down, gender; (**b**) hair, handbag, hat, up, upcolor, downcolor.

For the DukeMTMC-reID dataset, we list the accuracy of 10 attributes for comparison. "upcolor" has eight attributes: upblack, upwhite, upred, uppurple, upgray, upblue, upgreen, and upbrown. "downcolor" has seven attributes: downblack, downwhite, downred, downgray, downblue, downgreen, and downbrown. "Upcolor" and "Downcolor" in Figure 6b represent the average accuracy of these related attributes. It can be found from Figure 6 that, except for the attributes of "Backpack", "Bag", and "Top", the recognition rates of the attributes are greatly improved compared to the method proposed by Yin [7].



(a) Attribute recognition rate a (b) Attribute recognition rate b

Figure 6. Attribute recognition accuracy on DukeMTMC-reID dataset. (**a**) backpack, bag, boots, gender, handbag; (**b**) hat, shoes, top, upcolor, downcolor.

To better verify the performance of feature extraction in person re-identification with attributes, we discuss its cross-domain capabilities, as shown in Table 3.

Table 3. Experiment results in cross-domain dataset.

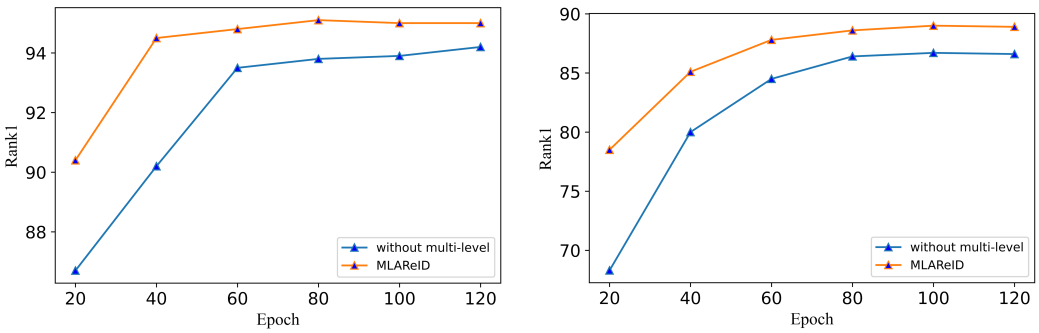
Method	M→D		D→M	
	Rank1	mAP	Rank1	mAP
TJ-AIDL(CVPR'18) [57]	44.3	23.0	58.2	26.5
SPGAN(CVPR'18) [58]	41.1	22.3	51.5	22.8
ATNet(CVPR'19) [59]	45.1	24.9	55.7	35.6
StrongReID [60]	41.4	25.7	54.3	25.5
SPGAN+LMP [58]	46.4	26.2	57.7	26.7
MLAReID	50.5	32.9	61.7	33.4
MLAReID + Reranking	55.4	46.7	65.6	48.2

It can be seen from Table 3 that the proposed method has advantages in Rank1 and mAP compared with other methods. M→D indicates that the source domain is Market-1501 and the target domain is DukeMTMC-reID, and D→M indicates the opposite. This verifies that the algorithm is effective in extracting pedestrian features after fully learning the relationship between pedestrian attribute labels and features, as well as between pedestrian identities and features.

4.5. Ablation Study

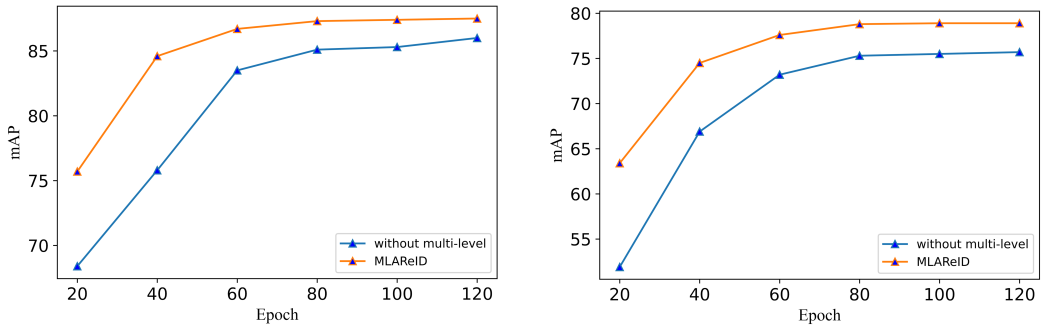
To better illustrate the effectiveness of the proposed method, we carried out ablation experiments for the three modules of non-local, instance batch normalization, and attributes. According to the Rank1, mAP, and mINP metrics, we evaluated whether to join the experimental results of these three modules.

It can be seen from Figures 7 and 8 that, compared with "without multi-level", both the performance of Rank-1 and mAP of MLAReID are improved more obviously in the iteration process, whether it be Market-1501 or DukemtMC-ReID. It shows that MLAReID proposed in this paper pays more attention to pedestrian images comprehensively and abstractly, and can better extract the semantic information of pedestrian images.



(a) Rank1 curve of Market-1501 (b) Rank1 curve of DukeMTMC-reID

Figure 7. Whether to add multi-level fusion module on different datasets (Rank1). Compared with without multi-level network, the proposed method achieves better Rank1 in training iteration.



(a) mAP curve of Market-1501 (b) mAP curve of DukeMTMC-reID

Figure 8. Whether to add multi-level fusion module on different datasets (mAP). Compared with without multi-level network, the proposed method achieves better mAP in training iteration.

In Table 4, "√" means to use a module, and a blank means to not use it. Without using the three modules of non-local, instance batch normalization, and attributes, we obtained 94.1% Rank1, 85.0% mAP, and 57.1% mINP on the Market-1501 dataset. When the attribute module was applied, Rank1 rose by 0.1%, mAP by 1%, and mINP by 2.1%. Application of the three modules produced much better results than the model with no modules. Rank1 improved by 2%, mAP by 5.3%, and mINP by 13.9%. For the DukeMTMC-reID dataset, without using the three modules, Rank1 was 85.9%, mAP was 74.8%, and mINP was 36.4%. After applying the three modules, Rank1 increased by 5.5%, mAP by 6.6%, and mINP by 11.5%. From the ablation experiment, we can see that these three modules improved network performance, and we verified the effectiveness of the proposed algorithm.

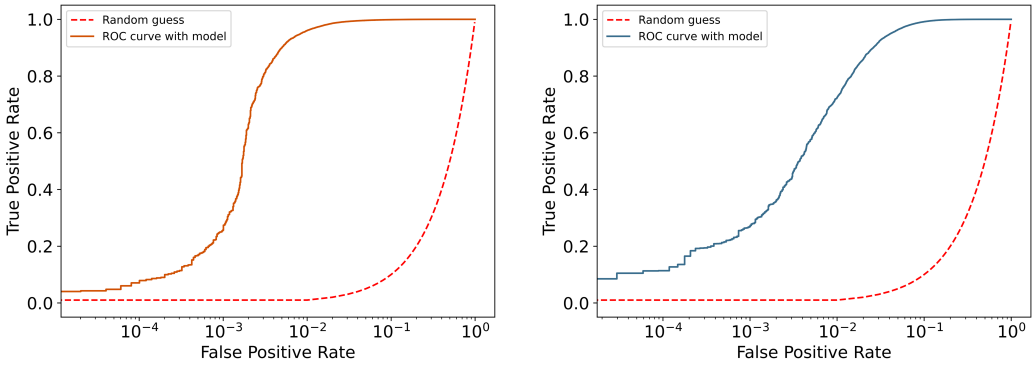
Table 4. Results of ablation study.

Component			Market-1501			DukeMTMC-reID		
Non-Local	IBN	Attribute	Rank1	mAP	mINP	Rank1	mAP	mINP
			94.1	85.0	57.1	85.9	74.8	36.4
		√	94.2	86.0	59.2	86.3	75.4	38.4
√		√	95.3	87.6	63.6	87.7	77.9	41.1
	√	√	96.0	89.5	69.2	90.5	80.2	45.2
√	√	√	96.1	90.3	71.0	91.4	81.4	47.9

In Figures 7 and 8, our ablation experiments on Market-1501 and DukeMTMC-reID for multi-level fusion module are listed. "MLAReID" represents a multi-level fusion model based on ResNet. "without multi-level" represents the model without a multi-level fusion module. As can be seen from Figures 7 and 8, for both Market-1501 and DukeMTMC-reID, the algorithm in this paper has a more comprehensive and abstract focus area for pedestrian images, which can better express the semantic information of images.

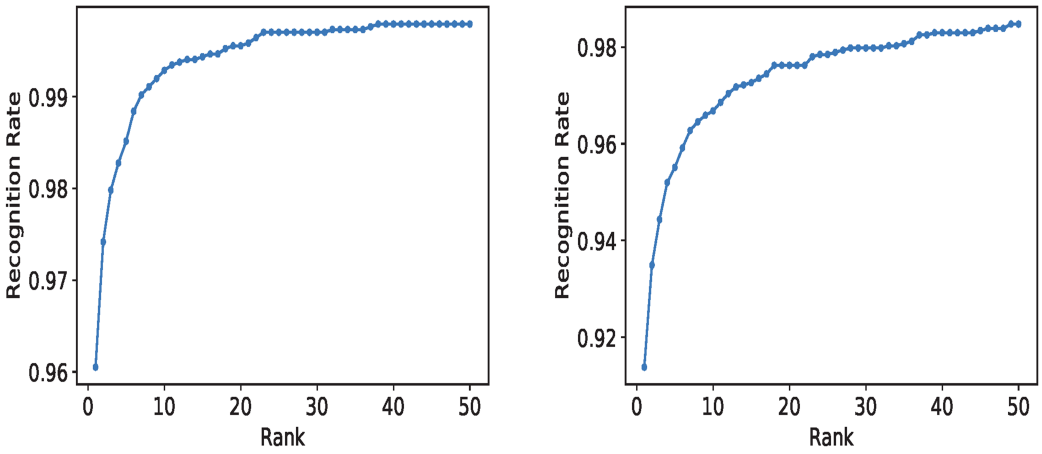
4.6. Visualization

We used a variety of visualization experiments to analyze the performance of the proposed method. Figures 9 and 10 show ROC and CMC curve on two datasets.



(a) ROC curve of Market-1501 (b) ROC curve of DukeMTMC-reID

Figure 9. Receiver operating characteristic (ROC) curve on different datasets.



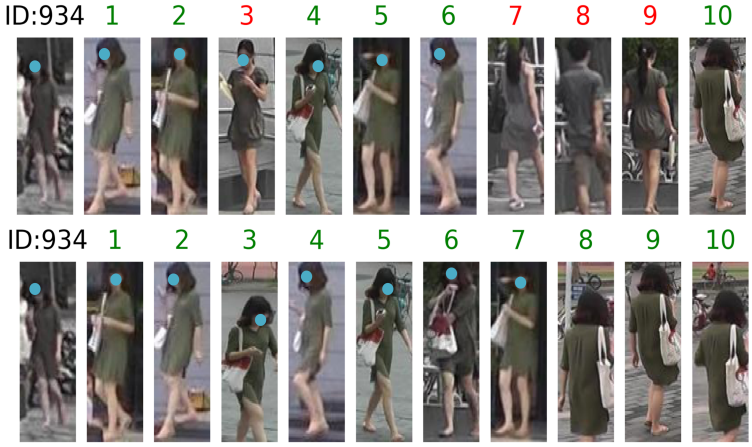
(a) CMC curve of Market-1501 (b) CMC curve of DukeMTMC-reID

Figure 10. Cumulative Match Curve (CMC) on different datasets.

To better verify the effectiveness of the proposed algorithm, we compared the visualization results of the two networks. "ID" in these figures (Figures 11–14) represents the query label, and serial numbers from 1 to 10 represent the sorting results from largest to smallest. Numbers in red indicate an incorrect match, and green indicates a correct match. The first line represents the visualization results of the baseline without attributes, and the second line represents the visualization results of the proposed method.

From Figure 11, we can find that the baseline (without local branch, non-local, IBN, and attribute) using no attributes has more mismatches. For example, for input image with ID 94, the top 10 images have many matching errors. This method makes an error in the "backpack" attribute, and regards a pedestrian without a backpack as an exact match. In addition, there is a matching error in the "clothing" attribute. For the pedestrian with ID 934, the baseline (without non-local, IBN, and attribute) does not correctly match the "hair" attribute. The method proposed in this paper accurately recognizes the key attributes of the pedestrian. The network has learned the relationship between key attributes of the pedestrian and pedestrian features, as well as between pedestrian identity labels and features.

(a)



(b)

Figure 11. Visualization results of networks (without local branch, non-local, IBN, and attribute); the second line is the result of the method proposed in this paper. We select two images with different IDs from the query. (**a**) Rank result visualization of ID 94; (**b**) Rank result visualization of ID 934.

From Figure 12, we selected pedestrian images with IDs 47 and 288 from the query. The baseline (without local branch, non-local, IBN, and attribute) that did not use attributes had more incorrect matches. For the pedestrian with ID 47, the baseline (without local branch, non-local, IBN, and attribute) had some errors on the "backpack" and "bag" attributes. For the pedestrian with ID 288, the baseline did not correctly match the "hair" attribute.

Figure 12. Comparison of visualization results of the two networks on DukeMTMC-reID. The first line is the visualization result of the baseline (without local branch, non-local, IBN, and attribute), and the second line is the result of the proposed method. We select two images with different IDs from the query. (**a**) Rank result of ID 47; (**b**) Rank result of ID 288.

We used GradCam to generate heat maps for the input pedestrians for comparison. For the two pedestrians from the Market-1501 dataset, the baseline (without local branch, non-local, IBN, and attribute) was compared with the proposed method. From Figure 13, we can find that the method of this paper focuses on parts-level features, and is more accurate than baseline. The proposed method accurately recognizes the key attributes of pedestrians, which demonstrates their important role in network parameter learning.

Figure 14 shows two examples for each dataset, one positive and one negative. Positive examples indicate that our proposed method can make correct predictions. Negative examples show that the proposed method can correctly predict attributes of pedestrians, but it is wrong on attributes that they do not possess. For example, for ID 94, our network predicts that the pedestrian has the "teenager", "backpack", "clothes", "up", "upblack", and "downblack" attributes, but ID 94 does not have the "downblue" attribute. For ID 329 of Market-1501 and ID 98 of DukeMTMC-reID, our network predicts them both completely accurately. The network not only predicts the attributes the pedestrian image has, but those the pedestrian image does not have.

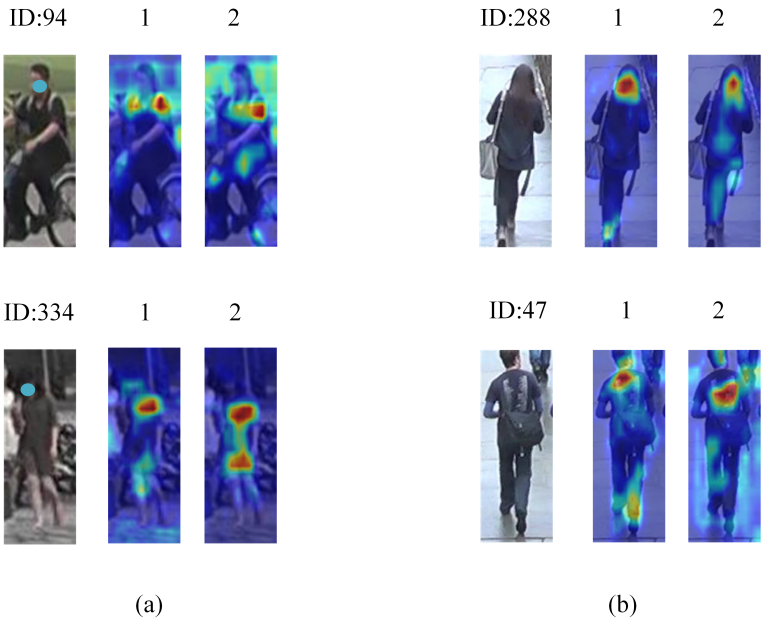


Figure 13. Comparison of heat map results of the two networks. "1" is the heat map result of the baseline (without local branch, non-local, IBN, and attribute); "2" is the heat map result of the method proposed in this paper. (**a**) ID 94, 934 on Market-1501; (**b**) ID 288, 47 on DukeMTMC-reID.

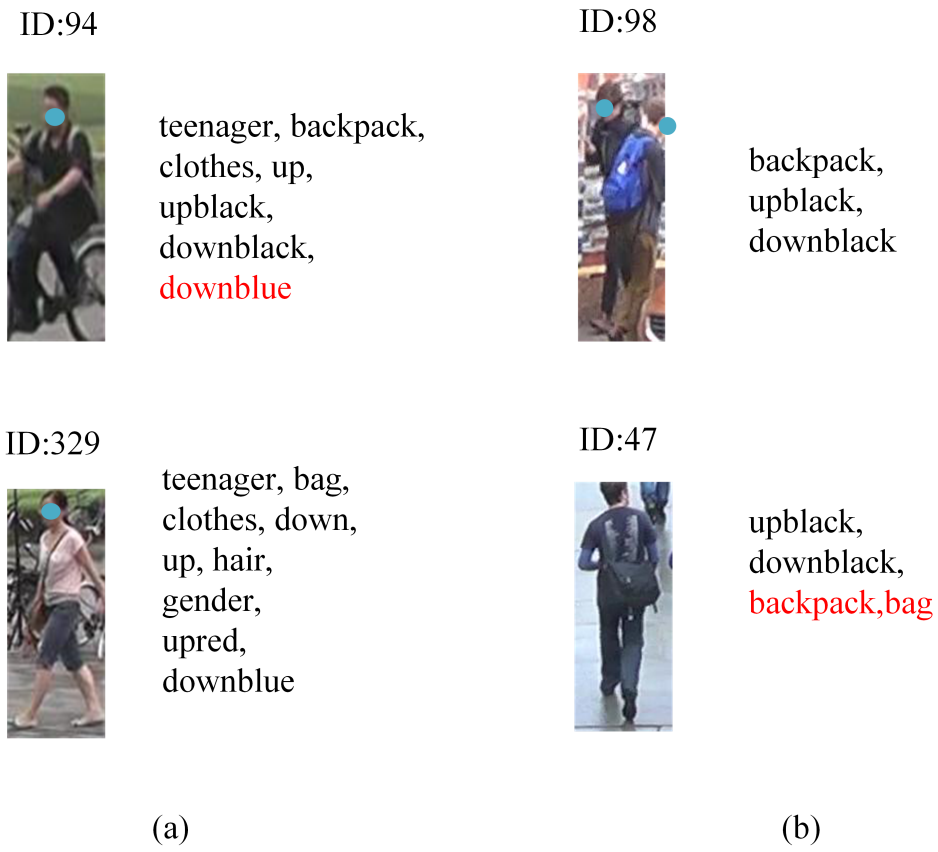


Figure 14. Attribute recognition results of proposed method. For each dataset, we list two examples, one positive and one negative. (**a**) Attribute recognition on Market-1501; (**b**) Attribute recognition on DukeMTMC-reID.

4.7. Time-Complexity Analysis

We provide the time complexity of the proposed method according to the network structure,

$$T_a = T[A_g(L4)] + N \times (T[C(E_g)] + T[L_a(C(E_g))]) \tag{10}$$

$$T_g = T[A_g(L4)] + T[C(E_g)] + T[L(E_g)] \tag{11}$$

$$T_i = T[A_l(L4)] + 6 \times (T[C(E_l)]) + T[L(E_l)] \tag{12}$$

$$T = T_a + T_g + T_l \tag{13}$$

where T_a, T_g, and T_l are the time complexities of the attribute, global, and local branch, respectively; T is their sum; $L4$ is the output of ResNet50; A_g is the aggregation function of the global branch; A_l is the aggregation function of the local branch; C is the classification function; E_g is feature embedding of the global branch; E_l is feature embedding of the local branch; and N is the number of attributes. The image is cut into six equal parts in this paper.

5. Conclusions

Each pedestrian has different attributes, and the number of attributes is variable, leading to unbalanced data distribution. In addition, pedestrians walk, and so the images captured by cameras are usually blurred, making it challenging to identify pedestrian attributes. This paper introduces a local information alignment module which focuses on specific regions. It combines a multi-task learning module and global module by learning the relationship between pedestrian attribute semantics and pedestrian identity. Our method solves the low attribute-recognition precision caused by unbalanced data distribution and blurred pedestrians to a certain extent. Through the experiments of two datasets, the multi-level network proposed in this paper can improve the precision of attribute recognition and enhance the performance of person Re-ID.

In the future, we will first generate images through the Generative Adversarial Network (GAN) to deal with unbalanced data distribution and low-quality pedestrian images. In addition, pedestrian images can be generated according to given attributes to increase training data. Second, we can introduce Graph Neural Network (GNN) and attention mechanisms to learn pedestrian posture information. We will focus on specific areas through alignment of pedestrian posture to identify pedestrian attributes more accurately.

Author Contributions: S.P. designed the algorithm, analyzed the experimental data, and wrote the manuscript. X.F. provided supervision, funding, and experimental equipment. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China under Grant Nos. 61402540, 60903222, 61672538, and 61272024, and in part by the Hunan Provincial Science and Technology Foundation under Grant No. 2014GK3049.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study can be found here: https://github.com/vana77/Market-1501_Attribute, https://github.com/vana77/DukeMTMC-attribute (accessed on 9 March 2022).

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

References

1. Bhattacharya, J.; Sharma, R.K. Ranking-based triplet loss function with intra-class mean and variance for fine-grained classification tasks. *Soft Comput.* **2020**, *24*, 15519–15528. [CrossRef]
2. Zhang, L.; Li, K.; Zhang, Y.; Qi, Y.; Yang, L. Adaptive image segmentation based on color clustering for person re-identification. *Soft Comput.* **2017**, *21*, 5729–5739. [CrossRef]
3. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *in press*. [CrossRef]
4. Yaghoubi, E.; Kumar, A.; Proena, H. Sss-pr: A short survey of surveys in person re-identification. *Pattern Recognit. Lett.* **2021**, *143*, 50–57. [CrossRef]
5. Wu, A.; Zheng, W.; Yu, H.; Gong, S.; Lai, J. RGB-infrared cross-modality person re-identification. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5390–5399.
6. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *arXiv* **2019**, arXiv:1703.07220.
7. Yin, J.; Fan, Z.; Chen, S.; Wang, Y. In-depth exploration of attribute information for person re-identification. *Appl. Intell.* **2020**, *50*, 3607–3622. [CrossRef]
8. Yaghoubi, E.; Khezeli, F.; Borza, D.; Kumar, S.V.A.; Neves, J.; Proena, H. Human Attribute Recognition—A Comprehensive Survey. *Appl. Sci.* **2020**, *10*, 5608. [CrossRef]
9. Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; Luo, B. Pedestrian attribute recognition: A survey. *Pattern Recognit.* **2022**, *121*, 108220. [CrossRef]
10. Han, K.; Huang, Y.; Song, C.; Wang, L.; Tan, T. Adaptive super-resolution for person re-identification with low-resolution images. *Pattern Recognit.* **2021**, *114*, 107682. [CrossRef]
11. Bai, S.; Li, Y.; Zhou, Y.; Li, Q.; Torr, P.H.S. Adversarial metric attack and defense for person re-identification. *IEEE Trans. Pattern Anal. Mach.* **2021**, *43*, 2119–2126. [CrossRef]
12. Zou, G.; Fu, G.; Peng, X.; Liu, Y.; Gao, M.; Liu, Z. Person re-identification based on metric learning: A survey. *Multimed. Tools Appl.* **2021**, *80*, 26855–26888. [CrossRef]
13. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv: 1610.02984.
14. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737
15. Sudowe, P.; Spitzer, H.; Leibe, B. Person attribute recognition with a jointly-trained holistic CNN model. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Washington, DC, USA, 7–13 December 2015; pp. 329–337.
16. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*; IEEE: Piscataway, NJ, USA, 2015; pp. 111–115.
17. Abdulnabi, A.H.; Wang, G.; Lu, J.; Jia, K. Multi-task CNN model for attribute prediction. *IEEE Trans. Multimed.* **2015**, *17*, 1949–1959. [CrossRef]
18. Zhu, J.; Liao, S.; Yi, D.; Lei, Z.; Li, S.Z. Multi-label CNN based pedestrian attribute learning for soft biometrics. In Proceedings of the 2015 International Conference on Biometrics (ICB), Phuket, Thailand, 19–22 May 2015; pp. 535–540.
19. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X. Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification. In Proceedings of the Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification, Long Beach, CA, USA, 16–20 June 2019; pp. 4913–4922.
20. Song, W.; Zheng, J.; Wu, Y.; Chen, C.; Liu, F. Partial attribute-driven video person re-identification. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; pp. 539–546,
21. Radenovi, F.; Tolias, G.; Chum, O. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1655–1668. [CrossRef]
22. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
23. Pan, X.; Luo, P.; Shi, J.; Tang, X. Two at once: Enhancing learning and generalization capacities via IBN-net. *arXiv* **2020**, arXiv:1807.09441.
24. Jin, H.; Lai, S.; Qian, X. Occlusion-sensitive Person Re-identification via Attribute-based Shift Attention. *IEEE Trans. Circ. Syst. Video Technol.* **2021**, *in press*. [CrossRef]
25. Xu, S.; Luo, L.; Hu, S. Attention-based model with attribute classification for cross-domain person re-identification. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9149–9155.
26. Taherkhani, F.; Dabouei, A.; Soleymani, S.; Dawson, J.; Nasrabadi, N.M. Attribute Guided Sparse Tensor-Based Model for Person Re-Identification. *arXiv* **2021**, arXiv:2108.04352.
27. Chen, X.; Liu, X.; Liu, W.; Zhang, Xi.; Zhang, Y.; Mei, T. Attrimeter: An attribute-guided metric interpreter for person re-identification. *arXiv* **2021**, arXiv:2103.01451.
28. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1116–1124.

29. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision—ECCV 2016 Workshops*; Springer: Berlin, Germany, 2016; Volume 9914, pp. 17–35.
30. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3774–3782.
31. Ustinova, E.; Ganin, Y.; Lempitsky, V. Multi-region bilinear convolutional neural networks for person re-identification. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
32. Jose, C.; Fleuret, F. Scalable metric learning via weighted approximate rank component analysis. In *Computer Vision—ECCV 2016*; Lecture Notes in Computer Science; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin, Germany, 2016; pp. 875–890.
33. Chen, D.; Yuan, Z.; Chen, B.; Zheng, N. Similarity learning with spatial constraints for person re-identification. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1268–1277.
34. Matsukawa, T.; Suzuki, E. Person re-identification using CNN features learned from combination of attributes. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2428–2433.
35. Varior, R.R.; Haloi, M.; Wang, G. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision 2016*; Springer: Amsterdam, The Netherlands, 8–16 October 2016; pp. 791–808.
36. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned CNN embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *14*, 13:1–13:20. [CrossRef]
37. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, 22–25 July 2017; pp. 7398–7407.
38. Zhao, L. *Deeply-Learned Part-Aligned Representations for Person Re-Identification*; In Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3239–3248.
39. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. *SVDNet for Pedestrian Retrieval*; In Proceedings of the International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 3800–3808.
40. Tian, M.; Yi, S.; Li, H.; Li, S.; Zhang, X.; Shi, J.; Yan, J.; Wang, X. Eliminating background-bias for robust person re-identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5794–5803.
41. He, L.; Liang, J.; Li, H.; Sun, Z. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7073–7082.
42. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2119–2128.
43. Qian, X.; Fu, Y.; Xiang, T.; Wang, W.; Qiu, J.; Wu, Y.; Jiang, Y.; Xue, X. *Pose-Normalized Image Generation for Person Re-Identification*; In *Computer Vision–ECCV 2018*; Springer: Berlin, Germany, 2018; pp. 650–667.
44. Lan, X.; Zhu, X.; Gong, S. Person search by multi-scale matching. In *Computer Vision–ECCV 2018*; Springer: Berlin, Germany, 2018; pp. 553–569.
45. Yu, R.; Dou, Z.; Bai, S.; Zhang, Z.; Xu, Y.; Bai, X. Hard-aware point-to-set deep metric for person re-identification. *arXiv* **2018**, arXiv:1807.11206.
46. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-aligned bilinear representations for person re-identification. In *Computer Vision–ECCV 2018*; Springer: Berlin, Germany, 2018; pp. 418–437.
47. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Computer Vision–ECCV 2018*; Springer: Berlin, Germany, 2018; pp. 501–518.
48. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 13–18 June 2018; pp. 420–429.
49. Yu, T.; Li, D.; Yang, Y.; Hospedales, T.M.; Xiang, T. *Robust Person Re-Identification by Modelling Feature Uncertainty*; In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 552–561.
50. Liu, Z.; Wang, J.; Gong, S.; Tao, D.; Lu, H. Deep reinforcement active learning for human-in-the-loop person re-identification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6121–6130.
51. Jiang, B.; Wang, X.; Tang, J. AttKGCN: Attribute knowledge graph convolutional network for person re-identification. *arXiv* **2019**, arXiv:1911.10544.
52. Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; Yang, A.Y. Salience-guided cascaded suppression network for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 3300–3310.

53. Zhao, C.; Tu, Y.; Lai, Z.; Shen, F.; Shen, H.T.; Miao, D. Salience-guided iterative asymmetric mutual hashing for fast person re-identification. *IEEE Trans. Image Process.* **2021**, *30*, 7776–7789. [CrossRef] [PubMed]
54. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Learning generalisable omni-scale representations for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *in press*. [CrossRef] [PubMed]
55. Li, Y.; Liu, L.; Zhu, L.; Zhang, H. Person re-identification based on multi-scale feature learning. *Knowl.-Based Syst.* **2021**, *228*, 107281. [CrossRef]
56. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
57. Wang, J.; Zhu, X.; Gong, S.; Li, W. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2275–2284.
58. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 994–1003.
59. Zhong, Z.; Zheng, L.; Li, S.; Yang, Y. Generalizing a Person Retrieval Model Hetero- and Homogeneously; In *European Conference on Computer Vision 2018*; Springer: Munich, Germany, 2018; pp. 172–188.
60. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. *IEEE Trans. Multimed.* **2019**, *22*, 2597–2609.