Multitask Model for Person Re-identification by Attribute-aware

Shengyu Pei (✉ shengyupei@csu.edu.cn)

Central South University <https://orcid.org/0000-0003-0924-3079>

Xiaoping Fan

Hunan University of Finance and Economics

Research Article

Keywords: Person re-identification , Attribute-aware , Non-local , Instance batch normalization

Posted Date: April 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-411924/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Multitask model for person re-identification by attribute-aware

Shengyu Pei · Xiaoping Fan

Received: date / Accepted: date

Abstract The image-based person re-identification problem can be transformed into a similar image retrieval problem. At present, most of the current identity-based methods do not consider pedestrian attributes. Moreover, many methods that consider pedestrian attributes and identities fail to fully simulate the relationship between pedestrian attributes and identities. In this article, we propose a new image-based person re-identification method by attribute-aware. Based on the introduction of instance batch normalization, the non-local module based on attention is used to transform the ResNet network structure to improve the feature extraction performance. After using generalized mean pooling for feature aggregation, the identity-based and attribute-based double stream network modules pay attention to the relationship between identities and pedestrian features, and the relationship between attributes and pedestrian features, so as to fully activate the relationship between pedestrian attributes and identities. Experiments are carried out on two classic person re-identification by attribute dataset Market-1501 and DukeMTMC-reID, and the results prove the effectiveness of the method. The method proposed in this paper has achieved the best performance on some evaluation metrics.

Keywords Person re-identification · Attribute-aware · Non-local · Instance batch normalization

Shengyu Pei
Central South University, School of Automation
E-mail: shengyupei@gmail.com

Xiaoping Fan
Hunan University of Finance and Economics, School of Information Technology and Management

1 Introduction

Given a pedestrian image, the purpose of person re-identification is to retrieve the images of the pedestrian from the cross-camera devices. It is generally believed that person re-recognition is an image retrieval technology designed to compensate for the visual limitations of fixed cameras. It can be combined with pedestrian detection and pedestrian tracking technologies to be used in industrial fields such as intelligent pedestrian detection and intelligent security [1, 2].

Due to the differences between camera devices, pedestrians have both rigid and flexible characteristics, and their appearance is also 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, the similarity of the same identity is large. Based on the supervised person re-identification problem, it has the labeled information. The labeled information is slightly rough, and pedestrian attributes with different identities also have many commonalities. There are also 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, there are three pedestrians with different identities, namely (x_1, y_1) , (x_2, y_2) and (x_3, y_3) . Suppose $y_1 \neq y_2 \neq y_3$, although x_1 and x_2 are very similar, but the similarity $S(x_1, x_2)$, $S(x_1, x_3)$ and $S(x_2, x_3)$ will be widened. As a result, the network pays attention to other regional feature information. When the 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. So we can draw a conclusion $S(x_1, x_2) > \max(S(x_1, x_3), S(x_2, x_3))$. 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. On the other hand, attributes also help the network speed up training. Adding attributes can filter out some pedestrian images that are not compatible with query attributes.

Of course, not that the more attributes the better the performance. In the image-based person re-recognition dataset, the two classic datasets that have been marked with attributes information include Market-1501 and DukeMTMC-reID. The Market-1501 dataset has 30 attributes for each identity, and the DukeMTMC-reID dataset has 23 attributes for each identity. There are many cases in the dataset where there are pedestrians with different identities, but the attributes of the pedestrians are very similar. That is, suppose two pedestrians with different identities, (x_1, y_1, A_1) and (x_2, y_2, A_2) , although there exists $A_1 = A_2$, but $y_1 \neq y_2$. It is not possible to directly use attribute labels

to train the network, but to choose discriminative attributes. Therefore, this paper chooses to use identity labels and attribute labels to train the network jointly. Person re-recognition with multi-modal attributes, although current single-modal algorithms have achieved good performance on some standard datasets. As we all know, RGB images are sensitive to bad weather such as rain, snow, and fog. However, intelligent monitoring needs to meet actual needs. The intuitive idea is to mine useful information from other modalities (such as thermal sensors, depth sensors, etc.), and then fuse them with RGB sensors. Therefore, there has been a lot of work deep fusion of multi-modal data to obtain performance improvement [3]. In recent years, some scholars have also proposed to introduce attributes to person re-identification tasks. Lin et al. [4] proposed an attribute person recognition (APR) network. It is a multi-task network that learns pedestrian ID embedding and predicts pedestrian attributes at the same time. 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. [5] proposed an Identity Recognition Network (IRN) and Attribute Recognition Network (ARN). Identity recognition network is used to extract partial information of pedestrian. Attribute recognition network is used to calculate the attribute similarity of pedestrians. Because of the role of attributes in detection and recognition, some scholars have also introduced them into video-based person re-recognition.

However, at present, person re-identification with attributes still encounters such problems.

1. At present, the accuracy of person re-identification with attributes is low. This paper proposes a new method to improve the accuracy of person re-identification with attributes.
2. The effect of person re-identification is very poor on cross-domain problems. The method proposed in this paper can improve the performance of cross-domain problems.
3. Compared with the latest person re-identification methods, the method proposed in this paper gets better performance.

2 Related work

Person re-identification solves the problem of matching pedestrian images between unrelated cameras. Its challenges are caused by different perspectives, postures, occlusion and other issues. To solve these problems, we need to increase the distance of inter-class and reduce the distance of intra-class. The current traditional methods include metric learning methods [6–8] and deep learning methods [9, 10].

Learning each attribute independently is a very 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 a task. Due to the high efficiency of multi-task learning, researchers have been paid more and more attention to it [11–14].

Lin et al. [4,5] proposed a network framework combining pedestrian ID labels and pedestrian attributes. These network frameworks break through the traditional limitation of learning only using pedestrian ID labels. By introducing pedestrian attribute labels, they construct a multi-task network that learns pedestrian ID labels and predicts pedestrian attribute labels at the same time. Because pedestrian is both flexible and rigid, and the camera devices are different, which greatly affected by the environment and the outside world, person re-identification is extremely challenging. These multi-task networks have high accuracy in attribute recognition, but lack reliable performance in person re-identification. Different from these networks, to improving the accuracy of person re-identification, this paper uses pedestrian attribute labels as an aid, and discusses the function of pedestrian attribute labels in pedestrian recognition.

Zhu et al. [14] used attributes to assist person re-identification network. They fuse low-level feature distance and attribute-based distance as the final distance to distinguish whether a given image is of the same identity.

Because of the role of attributes in detection and recognition, attributes have also been introduced in video-based person re-recognition. Zhao et al. [15] proposed an attribute-driven method for feature decomposition and frame weighting. The sub-features are re-weighted through the confidence of attribute recognition, and then integrated in the time dimension as the final representation. Through this strategy, the area with the largest amount of information in each frame is enhanced, and it contributes to a more differentiated sequence representation. Song et al. [16] proposed the partial attribute-driven network (PADNet). Current these methods are based on the global-level feature representation. Pedestrians are automatically divided into multiple body parts. A four-branch multi-label network is used to explore the spatio-temporal cues of the video.

The main work of person re-identification is based on static images. Although the ideas for solving the problem are different, the main idea is to transform the person re-identification problem into the most similar image retrieval problem. That is to say, 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, compare all pedestrian images in the gallery, and the image with closest distance is the identity of the query. Translated into the most similar image retrieval problem, 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 according to clothing, age, body, etc.

3 Proposed method

3.1 Non-local of residual network (ResNet) of instance batch normalization (IBN)

In this paper, we also added attention-like non-local module [17] and instance batch normalization (IBN) module [18] to learn more robust features.

The generalized non-local operation can be defined as:

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

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

Here, we apply non-local operations to ResNet of instance batch normalization. In order to adapt, we made the following changes.

$C(x) = N$, N represents the number of location of x . $f(x_i, x_j)$ uses dot product to calculate the affinity i and j . The formula is as follows:

$$f(x_i, x_j) = \theta(x_i)^T \phi(x_j) \quad (2)$$

In this paper, non-local block is acted on Layer 2, Layer 3 and Layer 4 of ResNet of instance batch normalization.

3.2 Optimization

In the person re-identification dataset, we can use $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ to represent the set with identity labels. Here, x_i represents the input image of the identity i . y_i represents the label of the identity i , and n represents the total number of input images. In the person re-identification dataset containing attributes, we can use A to represent the attribute set of all identities. Where $A_i = (A_i^1, A_i^2, \dots, A_i^m)$ is the attribute subset of the identity i , and m is the number of attributes of each identity. Therefore, we can use $E = \{(x_1, A_1), (x_2, A_2), \dots, (x_n, A_n)\}$ 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. Therefore, we can define the following three functions.

For the set D with identity labels, we define two functions, both of which are based on the objective function of identity labels. The first one is a classification function based on identity labels.

$$F_{Id} = \min_{w_{Id}, \theta} \sum_{i=1}^n L_{Id}(f_{Id}(w_{Id}, \phi(\theta, x_i)), y_i) \quad (3)$$

Here, $L_{Tri}(\phi, y_i)$ represents the metric loss function of the feature embedding identity. The purpose of the function F_{Tri} is to find the appropriate image feature embedding so that the identities of the same labels are as close as possible, and the identities of different labels are separated as much as possible.

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

$$F_{Att} = \min_{w_{Att}, \theta} \sum_{i=1}^n \sum_{j=1}^m L_{Att}(f_{Att_j}(w_{Att_j}, \phi(\theta, x_i)), A_i^j) \quad (4)$$

Here, $f_{Att_j}(w_{Att_j}, \phi)$ represents the classification function of the feature embedding the attribute of identity j . $L_{Att}(f_{Att_j}, A_i^j)$ indicates the j -th classification loss function of the identity i . Integrate all the attributes of the identity i together to get the attribute set of the identity i . 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 into the feature space. Then according to the Euclidean distance between each query image and all gallery images, the identity label of the query image is judged. $f_{Att_j}(w_{Att_j}, \phi(\theta, x_i))$ calculates all the attributes of each query image identity.

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

3.3 Network structure

In this section, we will describe the network structure in detail.

Figure 1 shows the network framework proposed in this paper. The network framework includes two parts: attribute recognition and identity recognition, corresponding to "A" and "B" of the framework diagram respectively. In the training phase, we first preprocesses the input images, and preprocessing plays the role of data enhancement. The preprocessing operation includes five different data enhancement modules: resize, random horizontal flip, pad, random crop and random erasing. The data enhancement modules can help to prevent the network from falling into local extrema during the training process, leading to overfitting.

It also helps realize the diversification of input images and helps the network better train. Then feed data into the ResNet backbone network used in this paper. In particular, this paper introduces two modules, instance batch normalization and non-local network, to reform ResNet so that ResNet has stronger feature extraction performance. The introduction of the two modules

Fig. 1 Network structure

of non-local and instance batch normalization has been introduced in section 2.1.

This paper introduces generalized mean pooling, its function is similar to adaptive average pooling. It is responsible for the aggregation of the feature embedding after the backbone. Its mathematical formula is as follows:

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

After feature aggregation, the aggregated features are sent to module A and module B respectively. Module A focuses on pedestrian attributes and is responsible for learning the relationship between pedestrian attributes and features. Module A can learn the correlation between each attribute of the pedestrian image and the features of the pedestrian image. Module B focuses on the identity of pedestrians, and is responsible for learning the relationship between pedestrian identity and features. Module B can learn the global features of pedestrian images. The combination of module A and module B makes the network pay attention to pedestrian attributes while also paying attention to the identity of pedestrians. It can prevent the phenomenon of network overfitting.

Fig. 2 Different batch normalization modules are used for different tasks. We design two modules to handle two different task. (a) Batch normalization (BN-1) module, for attribute recognition; (b) Batch normalization (BN-2) module, for identity recognition. IDs is the number of all different identities.

In module A, in order to obtain the relationship between each attribute and the aggregation features, a batch normalization (BN-1) module is set for each attribute. The BN-1 module is shown in the Figure 2 (a), followed by a two-classifier to determine whether the current pedestrian feature contains this attribute.

The "c" in module A represents a concat function, which refers to combining the classification results of all attributes. Finally, use BCE loss to calculate the difference between the predict attributes and the real attributes, and learn the relationship between pedestrian attributes and aggregated features through backpropagation.

In module B, in order to obtain the relationship between each identity and the aggregated features, a batch normalization (BN-2) module is also designed. This module has a similar function to the BN-1 in A, but only a simple one-dimensional batch normalization operation is used here. Finally, triplet loss, softmax loss and center loss are used to calculate the difference between the predict identity and the real identity. Through back propagation to learn the relationship between pedestrian identity and aggregation features, the network makes the pedestrians of the same identity as close as possible, and makes the pedestrians of different identities farther.

In the testing phase, there is a difference from the training phase. The network does not perform data enhancement on the input images, but in order to adapt to the network, it only performs a resize preprocessing 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, and find the images with the best rank score from the gallery through the distance matching method. "1", "2", "3" in module A and module B represent the output results of the network in the testing phase. The metrics in "2" are commonly used evaluation metrics for person re-identification problems. The embedded features in "1" are mainly used for output visualization. The metrics in "3" are commonly used evaluation metrics in person re-identification with attributes.

4 Experiment

In this section, we will conduct a lot of experiments to verify the effectiveness of the algorithm proposed in this paper. In order to distinguish, we call the method proposed in this paper as multi-task model for person re-identification by attribute-aware (MMAA).

4.1 Datasets and settings

1. Market-1501: The dataset is collected by 6 cameras in front of a supermarket in Tsinghua University. This dataset has 1501 identities and 32,668 annotated bounding boxes. Each annotated identity appears in at least 2 cameras. This 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 in the dataset has 30 different attributes. Note that although the upper body clothing and lower body clothing have 7 and 8 attributes, each identity has only one color marked as yes.
2. DukeMTMC-reID: The dataset comes from Duke University. This dataset contains 1,812 identities and 34,183 annotated bounding boxes. This dataset is divided into 702 training identities and 1,110 testing identities, corresponding to 16,522 and 17,661 images respectively. Attributes are annotated by pedestrian identity. Each image in the dataset has 23 different attributes.

4.2 Evaluation metrics

In order to measure the performance of the algorithm, this paper introduces standard metrics commonly used in person re-identification. 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, 4 different identities are fed into network in each batch, and each identity has 8 different images, that is, a total of 32 pedestrian images are fed into network in each batch.

4.4 Comparison with the state-of-the-art

In this section, this paper compares with the latest methods published in recent years. We choose Rank1, Rank5, Rank10 and mAP as the evaluation metrics.

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

Method	Rank1	Rank5	Rank10	mAP
MBC[19]	45.56	67	76	26.11
SML[20]	45.16	68.12	76	-
SL[21]	51.9	-	-	26.35
Attri[22]	58.84	-	-	33.04
S-CNN[23]	65.88	-	-	39.55
2Stream[24]	79.51	90.91	94.09	59.84
Cont-aware[25]	80.31	-	-	57.53
Part-align[26]	81.0	92.0	94.7	63.4
SVDNet[27]	82.3	92.3	95.2	62.1
GAN[28]	83.97	-	-	66.07
EBB[29]	81.2	94.6	97.0	-
DSR[30]	82.72	-	-	61.25
AACN[31]	85.90	-	-	66.87
APR[4]	87.04	95.10	96.42	66.89
PN-GAN[32]	89.4	-	-	72.6
CLSA[33]	88.9	-	-	73.1
HAP2S[34]	84.59	-	-	69.43
PABR[35]	90.2	96.1	97.4	76
PCB[36]	92.3	97.2	98.2	77.4
PSE[37]	87.7	94.5	96.8	69
DistributionNet[38]	87.26	94.74	96.73	70.82
DRAL[39]	84.2	94.27	96.59	66.26
AttKGCN[40]	94.4	98	98.7	85.5
Yin[5]	92.8	97.5	98.3	79.5
MMAA	96.1	98.5	99.3	90.3
MMAA+reranking	96.5	98.2	98.8	95.4

The following is a comparison result of Market-1501 and DukeMTMC-reID dataset. As shown in Table 1. In the table, "-" means no record.

For the Market-1501 dataset, the methods compared in this paper are GAN-based methods, part-based methods, and combined with attributes. From the Table 1, we find that compared with the current state-of-the-art methods, the results obtained in this paper achieve the best results on Rank1, Rank5, Rank10 and mAP.

The table lists the results obtained by the latest methods in recent years. In the Table 2, "-" means no record. For the DukeMTMC-reID dataset, compared with other methods on the evaluation metrics Rank1, Rank5, Rank10 and mAP, the method proposed in this paper achieve the best performance.

For the Market-1501 dataset, this paper lists the accuracy of 12 different attributes for comparison. Among them, "age" contains four different attributes, "young", "teenager", "adult", and "old". "upcolor" contains 8 different attributes: "upblack", "upwhite", "up-red", "uppurple", "upyellow", "upgray", "upblue", "upgreen". "downcolor" contains "downblack", "downwhite", "downpink", "downpurple", "downyellow", "downgray", "downblue", "downgreen", "downbrown" 9 different attributes. "upcolor" and "downcolor" respectively represents the average accuracy of these related attributes. It can be found from the Table 3 that "backpack", "bag", "hat", and "up" are not

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

Method	Rank1	Rank5	Rank10	mAP
BoW+kissme[41]	25.13	-	-	12.17
LOMO+XQDA[42]	30.75	-	-	17.04
AttrCombine[22]	53.87	-	-	33.35
GAN[28]	67.68	-	-	47.13
SVDNet[27]	76.7	-	-	56.8
APR[4]	73.92	-	-	55.56
PSE[37]	79.8	89.7	92.2	62
DistributionNet[38]	74.73	85.05	88.82	55.98
AttKGCN[40]	87.8	94.4	95.7	77.4
Yin[5]	82.7	91	93.5	66.4
MMAA	91.4	95.5	96.7	81.4
MMAA+reranking	92.7	96.1	97.2	90.6

Table 3 Attribute recognition accuracy on Market-1501

Method	Attribute age	backpack	bag	clothes	down	gender	hair	handbag	hat	up	upcolor	downcolor
Lin[4]	87.08	82.79	75.07	91.46	93.32	86.45	83.64	88.98	97.13	93.66	73.4	69.91
Yin[5]	87.7	90.2	81.9	94.9	94.7	91.5	89.1	90.5	97.5	93.8	76.4	68.6
MMAA	94.41	88.80	79.60	95.04	95.28	94.52	89.95	90.72	97.13	93.53	95.75	94.54

Table 4 Attribute recognition accuracy on DukeMTMC-reID

Method	Attribute backpack	bag	boots	gender	handbag	hat	shoes	top	upcolor	downcolor
Lin[4]	77.28	82.51	86.15	82.61	93.75	86.94	90.19	88	72.29	41.48
Yin[5]	85.4	83.4	89.9	85.9	93.4	88.7	91.6	89.5	75.5	70.1
MMAA	80.15	82.39	91.05	87.75	93.73	91.01	92.06	89.40	94.54	91.30

as good as the method proposed by Yin, but other attributes are significantly improved.

For the DukeMTMC-reID dataset, this paper lists the accuracy of 10 different attributes for comparison. Among them, "upcolor" contains 8 different attributes: "up-black", "upwhite", "upred", "uppurple", "upgray", "upblue", "upgreen", and "up-brown". "downcolor" contains 7 different attributes: "down-black", "downwhite", "downred", "downgray", "downblue", "downgreen", and "downbrown". "upcolor" and "downcolor" respectively represents the average accuracy of these related attributes. It can be found from the Table 4 that, except for the accuracy of the attributes of "backpack", "bag", and "top" compared with the method proposed by Yin, the other attributes are greatly improved compared to the current best method. In order to better verify the performance of feature extraction in person re-identification with attributes, we also discussed its cross-domain capabilities. As shown in the Table 5.

It can be seen from the Table 5 that the method proposed in this paper has advantages in Rank1 and mAP compared with other methods. This also

Table 5 Experiment results in cross-domain dataset

Method	Market-1501 $\rightarrow$ DukeMTMC-reID		DukeMTMC-reID $\rightarrow$ Market-1501	
	Rank1	mAP	Rank1	mAP
TJ-AIDL(CVPR'18)[43]	44.3	23.0	58.2	26.5
SPGAN(CVPR'18)[44]	41.1	22.3	51.5	22.8
ATNet(CVPR'19)[45]	45.1	24.9	55.7	35.6
StrongReID[46]	41.4	25.7	54.3	25.5
SPGAN+LMP[44]	46.4	26.2	57.7	26.7
MMAA	50.5	32.9	61.7	33.4
MMAA+Reranking	55.4	46.7	65.6	48.2

Table 6 Experiment 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

verifies that the algorithm is very effective in extracting pedestrian features after fully learning the relationship between pedestrian attribute labels and pedestrian features, as well as the relationship between pedestrian identities and pedestrian features.

4.5 Ablation study

In order to better illustrate the effectiveness of the method proposed in this paper, this paper carried out ablation experiments for the three modules of non-local, instance batch normalization, and attributes. According to Rank1, mAP and mINP metrics, evaluate whether to join the experimental results of these three modules.

From the Table 6, "✓" means using the corresponding module, and blank means not using the module. Without using the three modules of non-local, instance batch normalization and attributes, this paper obtained 94.1% Rank1, 85.0% mAP and 57.1% mINP in the Market-1501 dataset. When the attribute module is applied, Rank1 rises by 0.1%, mAP rises by 1%, and mINP rises by 2.1%. When the three modules are applied, it is much better than the model without any modules. Among them, Rank1 is improved by 2%, mAP is improved by 5.3%, and mINP is improved by 13.9%. For the DukeMTMC-reID dataset, without using the three modules of non-local, instance batch normalization and attributes, the Rank1 is 85.9%, mAP is 74.8%, and mINP is 36.4%. After applying the three modules, Rank1 increased by 5.5%, mAP increased by 6.6%, and mINP increased by 11.5%. From the ablation experiment, we

Fig. 3 A receiver operating characteristic (ROC) curve on different datasets. (a) ROC curve on Market-1501; (b) ROC curve on DukeMTMC-reID.

Fig. 4 Cumulative Match Curve (CMC) on different datasets. (a) CMC curve on Market-1501; (b) CMC curve on DukeMTMC-reID.

can see that these three models have improved network performance, and also verified the effectiveness of the algorithm proposed in this paper.

4.6 Visualization

In this section, we use a variety of visualization experiments to analyze the performance of the method proposed in this paper.

In order to better verify the effectiveness of the proposed algorithm, this paper compares the visualization results of the two networks. "ID" in these figures represents the query label, the serial number from 1 to 10 represents the sorting result from largest to smallest. Number in red font indicates an incorrect match, and number in green font 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 method proposed in this paper.

Fig. 5 Comparison of the visualization results of the two networks on Market-1501. The first line is the visualization result of the baseline, and the second line is the result of the method proposed in this paper. We select two images with different ID from the query. (a) Rank result of ID 94; (b) Rank result of ID 934.

Fig. 6 Comparison of the visualization results of the two networks on DukeMTMC-reID. The first line is the visualization result of the baseline, and the second line is the result of the method proposed in this paper. We select two images with different ID from the query. (a) Rank result of ID 47; (b) Rank result of ID 288.

From Figure 5, we can find that the baseline that do not use attributes have more mismatches. For example, input image with ID 94, the image of Rank1 matches wrongly. This method makes an error in "backpack" attribute, and regards pedestrian without backpack as an exact match. In addition, there is an error in matching in "clothing" attribute. For the pedestrian with ID 934, the baseline does not correctly match "hair" attribute. The method proposed in this paper accurately recognizes the key attributes of pedestrian. The network has learned the relationship between key attributes of pedestrian and pedestrian features, as well as the relationship between pedestrian identity labels and pedestrian features.

From Figure 6, we select pedestrian images with ID 47 and 288 from the query. The baseline that do not use attributes have more incorrectly matches. For the pedestrian with ID 47, the baseline has some errors on "backpack" and "bag" attributes. For the pedestrian with ID 288, the baseline does not correctly match "hair" attribute.

Fig. 7 Comparison of heat map results of the two networks. "1" is the heat map result of the baseline, "2" is the heat map result of the method proposed in this paper. (a) Heat map of ID 94, 934 on Market-1501; (b) Heat map of ID 288, 47 on DukeMTMC-reID.

We used GradCam to generate heat map for the input pedestrians for comparison. For the two pedestrians from the Market-1501 dataset, the baseline without attributes is used to compare with the method proposed in this paper. From Figure 7, we can find that the method mentioned in this paper focus on more parts and more accurate. The method proposed in this paper accurately recognizes the key attributes of pedestrians, and illustrates the important role of key attributes of pedestrians in network parameter learning.

From Figure 8, we listed two examples for each dataset, one positive and one negative. Positive examples indicate that our proposed method can predict correctly. The negative example shows that the method proposed in this paper can predict the attributes of pedestrians correctly, but it is wrong in attributes that they do not possess. For example, For ID 94, our network can predict that pedestrian contains "teenager", "backpack", "clothes", "up", "upblack" and "downblack", but ID 94 does not have the "downblue" attribute, but our network made a wrong judgment. For ID 329 of Market-1501 and ID 98 of DukeMTMC-reID, our network can predict completely accurately. The

(a) Attribute recognition on Market-1501 (b) Attribute recognition on DukeMTMC-reID

Fig. 8 Attribute recognition results of method proposed in this paper. For each dataset, we listed two examples, one positive and one negative. (a) Attribute recognition on Market-1501; (b) Attribute recognition on DukeMTMC-reID.

network not only predicts accurately the attributes that the pedestrian image has, but also predicts the attributes that the pedestrian image does not have.

5 Conclusions

Aiming at the low recognition accuracy of the current person re-identification by attribute and lack of discussion on cross-domain issues, this paper proposes a new person re-identification network with attribute-aware. The network introduces the instance batch normalization module, and uses the non-local method based on attention to transform the network. The method increases the performance of feature extraction. Bidirectional training of attribute-based networks and identity-based networks, the network can learn the relationship between attribute labels and pedestrian features, as well as the relationship between identity labels and pedestrian features. Our method is tested on two standard attribute person re-identification dataset, and the effectiveness of the method proposed in this paper is verified by comparing with the latest methods.

In the future, we will continue to carry out research from the following aspects:

1. We will apply reverse thinking and input pedestrian image attributes to filter unrelated pedestrian images, so as to achieve the purpose of fast pedestrian retrieval. Thereby providing convenience for commercial applications.
2. Use attention mechanism, graph neural network and other methods to further explore the correlation between attributes, and apply this idea to video-based person re-identification.

Compliance with ethical standards

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding details (In case of Funding)

This research was funded by the National Natural Science Foundation of China, grant number 61402540, 60903222, 61672538, and 61272024, and the Hunan Provincial Science and Technology Foundation, grant number 2014GK3049.

Conflict of interest

The authors declare that they have no conflict of interest.

Informed Consent

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2000.

Author's contribution

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Shengyu Pei, and Xiaoping Fan. The first draft of the manuscript was written by Shengyu Pei and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. J. Bhattacharya and R. K. Sharma. Ranking-based triplet loss function with intra-class mean and variance for fine-grained classification tasks. 24(20):15519–15528.
2. Lixia Zhang, Kangshun Li, Yan Zhang, Yu Qi, and Lei Yang. Adaptive image segmentation based on color clustering for person re-identification. 21(19):5729–5739.
3. A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai. RGB-infrared cross-modality person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5390–5399. ISSN: 2380-7504.
4. Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning.
5. Jianyuan Yin, Zheyi Fan, Shuni Chen, and Yilin Wang. In-depth exploration of attribute information for person re-identification. 50(11):3607–3622.
6. Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2288–2295. IEEE Computer Society.
7. S. Liao and S. Z. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3685–3693. ISSN: 2380-7504.
8. W. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. 35(3):653–668. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
9. Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. abs/1610.02984.
10. Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. abs/1703.07737. eprint: 1703.07737.
11. Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic CNN model. In *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, ICCVW '15, pages 329–337. IEEE Computer Society.
12. Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115. IEEE.
13. A. H. Abdalnabi, G. Wang, J. Lu, and K. Jia. Multi-task CNN model for attribute prediction. 17(11):1949–1959. Conference Name: IEEE Transactions on Multimedia.
14. J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *2015 International Conference on Biometrics (ICB)*, pages 535–540. ISSN: 2376-4201.
15. Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. pages 4913–4922.
16. W. Song, J. Zheng, Y. Wu, C. Chen, and F. Liu. Partial attribute-driven video person re-identification. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 539–546. ISSN: 2375-0197.
17. X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803. ISSN: 2575-7075.
18. Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via IBN-net.
19. E. Ustinova, Y. Ganin, and V. Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6.
20. Cijo Jose and François Fleuret. Scalable metric learning via weighted approximate rank component analysis. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 875–890. Springer International Publishing.

21. D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1268–1277. ISSN: 1063-6919.
22. T. Matsukawa and E. Suzuki. Person re-identification using CNN features learned from combination of attributes. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2428–2433.
23. Rahul Rama Vavior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - Eccv 2016, Pt Viii*, volume 9912, pages 791–808. WOS:000389500600048.
24. Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned CNN embedding for person reidentification. 14(1):13:1–13:20.
25. D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7398–7407. ISSN: 1063-6919.
26. Liming Zhao. Deeply-learned part-aligned representations for person re-identification. pages 3239–3248.
27. Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. SVDNet for pedestrian retrieval. pages 3800–3808.
28. Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3774–3782. ISSN: 2380-7504.
29. M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang. Eliminating background-bias for robust person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5794–5803. ISSN: 2575-7075.
30. L. He, J. Liang, H. Li, and Z. Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7073–7082. ISSN: 2575-7075.
31. J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. Attention-aware compositional network for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2119–2128. ISSN: 2575-7075.
32. Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. pages 650–667.
33. Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018, Lecture Notes in Computer Science*, pages 553–569. Springer International Publishing.
34. Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification.
35. Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018, Lecture Notes in Computer Science*, pages 418–437. Springer International Publishing.
36. Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018, Lecture Notes in Computer Science*, pages 501–518. Springer International Publishing.
37. M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 420–429. IEEE.
38. Tianyuan Yu, Da Li, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. pages 552–561.
39. Z. Liu, J. Wang, S. Gong, D. Tao, and H. Lu. Deep reinforcement active learning for human-in-the-loop person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6121–6130. ISSN: 2380-7504.

40. B. Jiang, Xixi Wang, and Jin Tang. AttKGCN: Attribute knowledge graph convolutional network for person re-identification.
41. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124. ISSN: 2380-7504.
42. S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206. ISSN: 1063-6919.
43. Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2275–2284. IEEE Computer Society.
44. Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 994–1003. IEEE.
45. Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. pages 172–188.
46. Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification.

Figures

Figure 1

Network structure

(a) Batch normalization (BN-1)

(b) Batch normalization (BN-2)

Figure 2

Different batch normalization modules are used for different tasks. We design two modules to handle two different task. (a) Batch normalization (BN-1) module, for attribute recognition; (b) Batch normalization (BN-2) module, for identity recognition. IDs is the number of all different identities.

Figure 3

A receiver operating characteristic (ROC) curve on different datasets. (a) ROC curve on Market-1501; (b) ROC curve on DukeMTMC-reID.

Figure 4

Cumulative Match Curve (CMC) on different datasets. (a) CMC curve on Market-1501; (b) CMC curve on DukeMTMC-reID.

(a) Rank result of ID 94

(b) Rank result of ID 934

Figure 5

Comparison of the visualization results of the two networks on Market-1501. The first line is the visualization result of the baseline, and the second line is the result of the method proposed in this paper. We select two images with different ID from the query. (a) Rank result of ID 94; (b) Rank result of ID 934.

(a) Rank result of ID 47

(b) Rank result of ID 288

Figure 6

Comparison of the visualization results of the two networks on DukeMTMC-reID. The first line is the visualization result of the baseline, and the second line is the result of the method proposed in this paper. We select two images with different ID from the query. (a) Rank result of ID 47; (b) Rank result of ID 288.

(a) Heat map on Market-1501

(b) Heat map on DukeMTMC-reID

Figure 7

Comparison of heat map results of the two networks. "1" is the heat map result of the baseline, "2" is the heat map result of the method proposed in this paper. (a) Heat map of ID 94, 934 on Market-1501; (b) Heat map of ID 288, 47 on DukeMTMC-reID.

ID:94

teenager, backpack,
clothes, up, upblack,
downblack, **downblue**

ID:98

backpack, upblack,
downblack

ID:329

teenager, bag,
clothes, down,
up, hair, gender,
upred, downblue

ID:47

upblack,
downblack,
backpack, bag

(a) Attribute recognition on Market-1501

(b) Attribute recognition on DukeMTMC-reID

Figure 8

Attribute recognition results of method proposed in this paper. For each dataset, we listed two examples, one positive and one negative. (a) Attribute recognition on Market-1501; (b) Attribute recognition on DukeMTMC-reID