LRDNN: Local-refining based Deep Neural Network for Person Re-Identification with Attribute Discerning

Qinqin Zhou[1], **Bineng Zhong**[1,2*], **Xiangyuan Lan**[3*], **Gan Sun**[4,5], **Yulun Zhang**[6], **Mengran Gou**[6]
[1]Department of Computer Science and Technology, Huaqiao University
[2]Provincial Key Laboratory for Computer Information Processing Technology, Soochow University
[3]Department of Computer Science, Hong Kong Baptist University
[4]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences
[5]University of Chinese Academy of Sciences
[6]Department of ECE, Northeastern University
{qinqinzhou, bnzhong}@hqu.edu.cn, xiangyuanlan@life.hkbu.edu.hk,
{sungan1412,yulun100}@gmail.com, mgou@ieee.org

Abstract

Recently, pose or attribute information has been widely used to solve person re-identification (re-ID) problem. However, the inaccurate output from pose or attribute modules will impair the final person re-ID performance. Since re-ID, pose estimation and attribute recognition are all based on the person appearance information, we propose a Local-refining based Deep Neural Network (LRDNN) to aggregate pose estimation and attribute recognition to improve the re-ID performance. To this end, we add a pose branch to extract the local spatial information and optimize the whole network on both person identity and attribute objectives. To diminish the negative affect from unstable pose estimation, a novel structure called channel parse block (CPB) is introduced to learn weights on different feature channels in pose branch. Then two branches are combined with compact bilinear pooling. Experimental results on Market1501 and DukeMTMC-reid datasets illustrate the effectiveness of the proposed method.

1 Introduction

Nowadays, it's impractical to manually analyze massive surveillance videos from a large camera network. Person re-identification (re-ID) is proposed to mitigate the human effort by processing the cross-camera person recognition task efficiently. Specifically, given a target image as the probe instance, re-ID aims to find the same person in a set of gallery images from different cameras. Although several large benchmarks [Zheng *et al.*, 2015; Zheng *et al.*, 2017b] have been introduced, re-ID problem is still far from being solved. In the past decade, many works [Sarfraz *et al.*, 2018; Suh *et al.*, 2018] try to address the issues of re-ID based on the global feature from the entire image which may be biased by the background information and body parts misalignment. Recently, [Liu *et al.*, 2017] adopts the attention mechanism to focus on local areas that are in favor of attribute recognition. [Li *et al.*, 2014] refines the matching by dividing the human body into different regions. In [Zhao *et al.*, 2017a], pose estimation method is applied to locate body parts and then a normalized person image will be used for re-ID. However, artificially pre-defined division and combination strategies for body parts also introduce the background information.

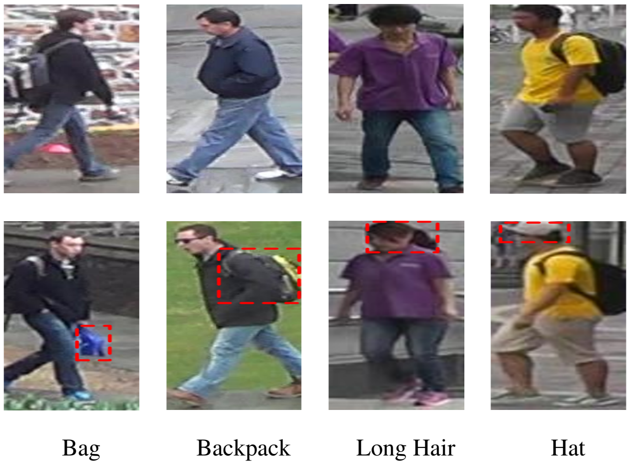
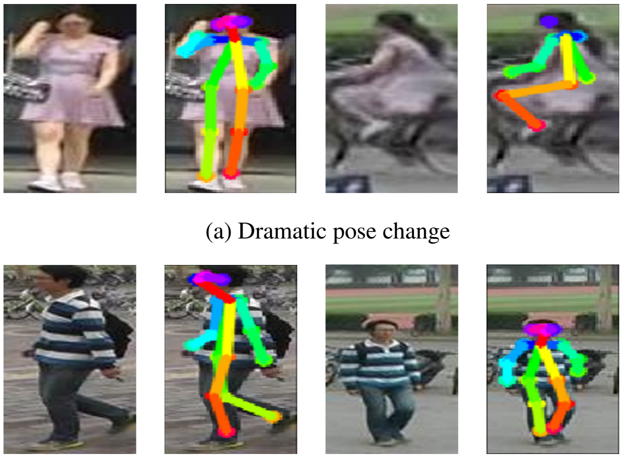


Figure 1: Attributes in person re-identification. Each column shows two different people in similar appearance, while they can be easily distinguished by attributes (e.g. carrying bag, carrying backpack, the length of hair, and wearing hat).

In this paper, we propose a simple yet effective framework to combine both pose estimation and attribute recognition onto the re-ID task. By adding a pose branch, the proposed local-refining based deep neural network (LRDNN) can extract the spatial information for the local body parts. After combining the pose branch and main branch, the attribute recognition objective will be optimized simultaneously to boost the re-ID performance. Furthermore, considering that

[*]Corresponding Author


(a) Dramatic pose change

(b) Redundant background

Figure 2: The examples of various poses in the re-ID scenes, where each row shows the different poses of one person. The original images are shown in the first and third columns. The pose information obtained by our model are shown in the second and fourth columns.

the different channels in the feature map has different characteristics, we also design a novel channel parse block (CPB) to automatically weight the channels in pose branch such that the weighted feature can reduce both re-ID and attribute losses. The pose branch is pre-trained on MSCOCO [Lin *et al.*, 2014] dataset to locate different body parts.

To sum up, the contributions of this work are three-fold:

- We propose a local-refining based deep neural network for person re-identification which composes of a main branch and a pose branch to fuse pose and attribute information in a consistent way. To the best of our knowledge, this is the first work which explicitly exploits the pose and attribute jointly.
- We design a simple yet effective pose driven method to simultaneously optimize attribute and re-ID losses. Consequently, our method can achieve promising performance on two large re-ID datasets.
- We design a robust structure to pick the most important body parts for re-ID, and thus can greatly reduce the errors caused by inaccurate pose estimation.

2 Related Work

Recently, many deep learning based person re-ID approaches with attribute and pose clue have been proposed which achieve promising results on many datasets. In this section, we will make a detailed introduction about the development of re-ID combined with the application of attribute clue and pose clue in the field of re-ID.

Person Re-ID: In the past literature, most of re-ID methods [Lin *et al.*, 2017; Zheng *et al.*, 2018] focus on global feature and do not distinguish between different local features. There are several works [Johnson *et al.*, 2018; Sun *et al.*, 2018] which present some feasible solutions to ease the misalignment problem. [Johnson *et al.*, 2018] uses detected local body regions to learn the reconstructed global deep feature and fuse handcrafted features to adjust the misalignment. In [Sun *et al.*, 2018], a network called Part-based Convolutional Baseline (PCB) is proposed to generate part-level features, then use a refined part pooling to adjust local distribution. However, these methods doesn't consider background impact. In recent years, deep learning based re-ID dominate the literature and has been applied in other areas. [Zheng *et al.*, 2016] uses deep convolution networks as their baseline to train models robust to illumination variations and cluttered background. [Zhou *et al.*, 2018] combines re-ID into the tracking framework to improve performance. Our method is inspired by the latest pose estimation algorithm [Cao *et al.*, 2018] and the main insight is to combine local parts of images, i.e. different body parts and human attributes with global feature to improve the performance of re-ID.

Re-ID with Attributes Recognition: Based on the attributes labeled by [Schumann and Stiefelhagen, 2017] on Market1501 [Zheng *et al.*, 2015] and DukeMTMC-reid [Zheng *et al.*, 2017b], adopting attribute information in re-ID becomes an emerging direction in recent years. As shown in Figure 1, like human tends to recognize an object with distinguishable markers, attribute plays an important role in re-ID. [Feng *et al.*, 2018] designs a deep network model with attribute learning which gains significant performance improvement on both Market1501 and DukeMTMC-reid. [Su *et al.*, 2016] use a semi-supervised deep model to perform attribute learning based on global features. Considering attribute recognition is based on both global and local appearance feature and a simple network cannot balance both sides perfectly, our method incorporates a main branch and a pose branch, and the pose branch produces pose clues to guide the main branch to recognize both human attributes and identity. Moreover, the feature maps fused by the pose branch and the main branch encodes fine-grained semantic information which can boost the performance on both tasks.

Re-ID with Pose Parsing: Since human is the primary research object of re-ID, there are many works [Sarfraz *et al.*, 2018] taking into account the nature of non-rigid human body and introducing strategies to merge the pose clue. As shown in Figure 2, human pose change and redundant background are common in re-ID scene and pose clue is useful for these situation. [Kalayeh *et al.*, 2018] proposes a segmentation based method to distinguish different parts of the human body and group those parts in a consistent way to prevent misalignment. [Zheng *et al.*, 2017a] designs a multi-branch deep neural network to capture the features of different human body parts and a body region proposal network to generate 10 body regions. While these algorithms still suffer from the loss of finer-grained pose information, we propose a method which encodes both a pixel based pose clue and attribute clue in a unified framework.

3 Our Method

In this paper, we propose a local-refining based deep neural network (LRDNN) which focus on semantic clue as well as refined local clue. More specifically, LRDNN includes two

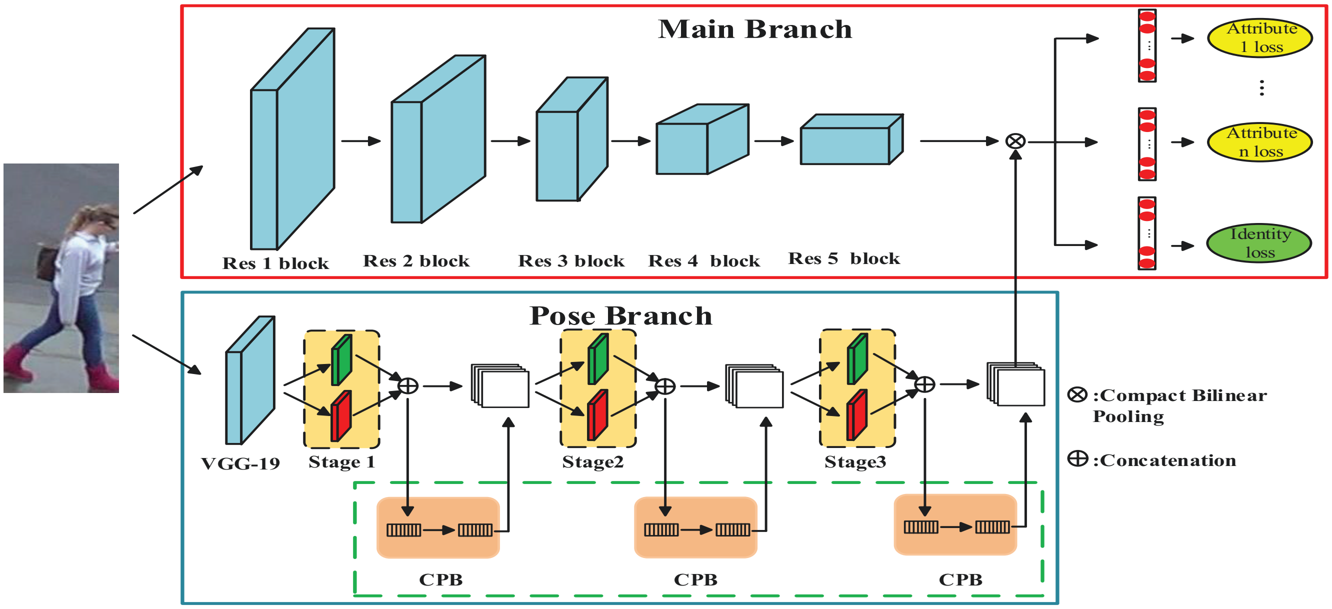


Figure 3: The framework of our LRDNN.

branches: the main branch M which encodes semantic cues with attribute information and the pose branch P which encodes local clue of human body. To reduce the influences caused by unstable pose estimation and occlusions, we introduce the channel parse block (CPB) to the pose branch which gives greater weights to more effective feature channels. With the pose clue produced by the pose branch, identity recognition and attribute discerning tasks learn to place more attention on human body. The integration of these three aspects can promote ID recognition and attribute discerning tasks performance. In the following subsections, we first introduce the framework of LRDNN. Then, we describe how the attribute clue and pose clue are used in our method. Finally, how CPB weights different channels of feature is presented in detail.

3.1 Local-refining based Deep Neural Network (LRDNN)

Although a significant progress has been achieved in deep learning-based re-ID, it is still a challenging problem for the issues of misalignment across cameras and redundant background. Meanwhile, fine-grained clues such as attributes and pose have received great attention from many researchers and recent works [Johnson *et al.*, 2018; Li *et al.*, 2014] proved its effectiveness. Since the redundant background and occlusions make the overall visual information become contaminated, incorporating attribute and pose clue into the re-ID framework can suppress the error caused by misalignment effectively. In addition, for similar people, discriminate attributes can help to distinguish them and the pose estimation not only provides useful local information for attribute recognition but also helps to align the body parts in different cameras. Our LRDNN contains two branches to depict attribute and pose clues with the deep semantic feature.

As shown in Figure 3, the main branch and the pose branch compose the chief part of LRDNN. There are two tasks: 1) identity recognition, 2) attribute discerning. Given an input image I to the main branch and the pose branch, the main branch encodes the semantic feature and the attributes clue of different identities which is expressed as:

$$\mathcal{F} = \mathbb{M}(I). \tag{1}$$

Correspondingly, based on [Cao *et al.*, 2018], the pose branch learns the part confidence maps and part affinity fields of local area on human body which are produced by the fourth stage.

$$(\mathcal{H}^3, \mathcal{C}^3) = \mathbb{P}(I), \tag{2}$$

where $\mathcal{H}^3$ and $\mathcal{C}^3$ denote the part confidence map and the part affinity field obtained in the fourth stage. Then, two branches are aggregated by compact bilinear pooling which can further highlight the human body and suppress irrelevant background information. To ease the instability of pose branch, we introduce the CPB to parse the channels of feature maps outputted by the pose branch.

3.2 Pose Clue

Pose clue plays an important role in computer vision tasks like re-ID, pose style transformation and abnormal behavior detection. Specifically, recent studies [Kalayeh *et al.*, 2018; Zheng *et al.*, 2017a] introduce the pose clue into re-ID task and significantly improve the performance. In this paper, we focus on finer grained pose clues which are specific to the joint points of human bodies. Inspired by the recently proposed pose estimation algorithm [Cao *et al.*, 2018] which uses a two-stream multi-stage CNN model to detect the key parts of human body and give the corresponding part affinity fields of these key parts simultaneously, we retrain a four-stage model as the main part of our pose branch. The structure of our pose branch is showed in Figure 3.

Datasets	gender	hair	L.slv	L.low	L.up	T.low	hat	B.pack	bag	H.bag	age	C.up	C.low	C.shoes	boots
Market1501	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		
Duke	✓				✓		✓	✓		✓	✓	✓	✓	✓	✓

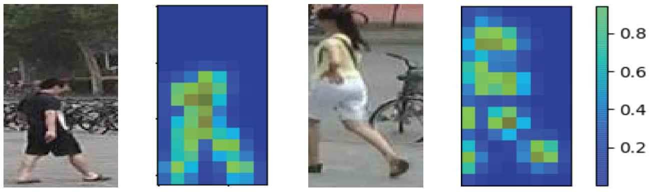
Table 1: Table of different attributes in different datasets. L.slv, L.low, L.up, B.pack, H.bag, C.up, C.low, C.shoes denote length of sleeve, length of lower-body clothing, length of upper-body clothing, backpack, handbag, color of upper-body clothing, color of lower-body clothing, and color of shoes, respectively.

In our framework, the task of the pose branch is to capture the key parts of human body and learn channel-based parsing rule. More specifically, the two streams S_1 and S_2 in the pose branch provide part confidence maps and part affinity fields of the key parts separately. For the 19 joints, it will produce 19 part confidence maps and 38 part affinity fields. Specifically, following [Cao *et al.*, 2018], the pose branch is composed of 4 stages which include one VGG and 3 designed streams pretrained on MSCOCO dataset. Given an input image I, we denote the output part confidence maps and part affinity fields of the k stage as H^k and C^k, respectively. It should be noted that the two streams in the pose branch are trained with the MSCOCO dataset first. More specifically, the pose branch are trained with two MSE (Mean Squared Error) loss functions for the two streams. Given the ground-truth maps $G_1^u(p)$, $G_2^v(p)$ and the estimated predictions $H_u^t(p)$, $C_v^t(p)$, the MSE losses for the two streams formulated as follows:

$$\mathcal{L}_{S_1}^t = \sum_{u=1}^{U} \sum_{p} W(p) \cdot \left\| H_u^t(p) - G_1^u(p) \right\|_2^2 , \tag{3}$$

$$\mathcal{L}_{S_2}^t = \sum_{v=1}^{V} \sum_{p} W(p) \cdot \left\| C_v^t(p) - G_2^v(p) \right\|_2^2 . \tag{4}$$

To make full use of pose clues in pose branch, we aggregate the combination of the part confidence maps and part affinity fields of pose branch with the feature maps of the main branch by compact bilinear pooling. More specifically, we add a batch normalization layer after the pose branch to prevent model overfitting and enforce the feature maps produced by these two branches to be subjected to the same distribution. As shown in Figure 4, after the pose clues are fused into our framework, the redundant background is suppressed effectively and weights are mainly concentrated on human body. Furthermore, the problem of inaccurate detection is quite common for person re-ID datasets which causes detection failures of the pose branch, so we proposed a channel parse block (CPB) to ease them. The details of the CPB are presented in Section 3.4.



(a) (b)

Figure 4: The visualization results after joining the pose clues. Both of (a) and (b) include redundant background.

3.3 Attribute Clue

Various attributes of human body have good properties in identity recognition task. Previous works have introduced attributes to re-ID as an enhanced descriptor for different persons. Due to the lack of datasets and in-depth analysis, [Wang *et al.*, 2017] just focus on simple attributes detection and do not have consider high-level tasks like re-ID. With the explosive development of deep learning algorithms, re-ID begins to make extensive use of deep learning technology and large person re-ID datasets with attribute labeled (e.g., Market1501, DukeMTMC-reid) are also available. These inspire us to combine attributes to improve person re-ID task.

In our framework, we choose ResNet-50 as the backbone network of the main branch and pre-trained weights on the ImageNet are used for better convergence. In the main branch of our framework, we add attribute discerning task to enhance the expression ability of semantic information for the identities. In order to adapt to our re-identification task, we replace the FC layer at the end of ResNet-50 with N FC classification layers for attributes recognition plus a FC layer for identity recognition. The attributes used in our framework on Market1501 and DukeMTMC-reid are listed in Table 1.

In our implementation, we treat the identity recognition and attribute discerning tasks as classification tasks. An image I is given to the main branch to encode the semantic feature $\mathcal{F}$ of the identity. To make the features from the main branch and the pose branch follow the same distribution, we normalized the outputs of two branches by using batch normalization as follows:

$$\mathcal{F} = \frac{f^{(k)} - \mu_f}{\sqrt{\sigma_f^2 + \varepsilon}}, \tag{5}$$

$$\mathcal{H}^3 = \frac{h^{(k)} - \mu_h}{\sqrt{\sigma_h^2 + \varepsilon}}, \tag{6}$$

$$\mathcal{C}^3 = \frac{c^{(k)} - \mu_c}{\sqrt{\sigma_c^2 + \varepsilon}}, \tag{7}$$

where μ_* and σ_* denote the corresponding mean and variance of a mini batch, and ε is a parameter to prevent the denominator from 0. After normalization, these three feature maps can be constrained in the same magnitude.

Then, we use compact bilinear pooling layer to aggregate these two branches. As shown in Figure 4, the features produced by the compact bilinear pooling layer can preserve the expression ability of two branches with much less memory

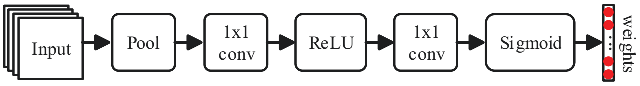


Figure 5: The detail structure of CPB.

cost. For the identity and the i-th attribute, we use softmax loss to train the network. Let $z_i = [z_1, z_2, \cdots, z_n]$ be the vector produced by the i-th FC layer, we calculate the softmax loss of the i-th attribute/identity as follow:

$$\mathcal{L}_M^q = -\sum_{N}^{n=1} log(p(n))q(n), \qquad (8)$$

where $p(n) = \frac{exp(z_n)}{\sum_{i=1}^{N} exp(z_i)}$ denotes the predicted possibility of each attribute or ID label. When the predicted result is same as the ground-truth, $q(n) = 1$.

3.4 Channel Parse Block(CPB)

Since the misalignment issue caused by human poses variations limits the performance of re-ID, many methods begin to use spatial segmentation with weak generalization ability [Li *et al.*, 2014; Zhao *et al.*, 2017a]. However, these methods focus on the spatial relationship between human body parts and ignore the unstable detection. Different from previous methods, we not only consider the spatial relationship but also take feature channel characteristics into account.

Our CPB model is inspired by the work of [Hu *et al.*, 2017] from the image recognition community, which presents an architecture called SE-net block to weight different channels of features by modelling inter-dependencies between channels explicitly. To unify the two-branch model and the CPB model in the learning framework, we modify the FC layer in SE-net to 1x1 convolutional layer as our channel parse block. The detailed structure is shown in Figure 5. As shown in Figure 3, we aggregate the CPB model with our pose branch and train it to parse different channels of feature maps. More specifically, the different channels of feature maps in our pose branch characterize different specific parts of the human body, which makes CPB not only learn to assign large weights to those channels in favor of identity recognition and attribute discerning, but also reduce the estimation inaccuracy in pose branch.

4 Experiments

In this section, we present the experimental resuls of our method on Market1501 and DukeMTMC-reid datasets and compare the results with the state-of-art methods. To illustrate the effectiveness of the different components in our method, well-designed ablation analysis is also presented.

4.1 Datasets

Market1501[Zheng *et al.*, 2015]: This dataset includes 1,501 identities captured by 6 different cameras. More specifically, there are 12,936 training images, 19,732 testing images and 3,368 query images detected by the DPM model [Felzenszwalb *et al.*, 2010]. These images are with the same resolution of 128 × 64, among which the training images and the testing images contain 751 and 750 identities, respectively.

DukeMTMC-reid[Zheng *et al.*, 2017b]: DukeMTMC-reid is another large dataset which is used in multi-target tracking and person re-ID. It contains 16,522 training images, 17,661 testing images and 2,228 query images which cover 1,812 identities. All these images are captured at the Duke University campus by 8 synchronized cameras, then annotated and aligned manually.

4.2 Evaluation Protocol

In our experiments, we adopt the common evaluation protocols in person re-ID which include the cumulative matching characteristic (CMC) curve and the mean average precision (mAP) for Market1501 and DukeMTMC-reid. The CMC curve is a precision curve that provides recognition precision for each rank. As a supplement, the mAP is the metric to measure the accuracy of person re-ID and it is the average of the maximum precisions at different recall values.

4.3 Implementation Detail

In our implementation, the person re-ID model proposed in this paper is constructed in tensorflow. The configuration of the experimental environment is a HP Pro 3330 PC with Intel 1.8GHz×32 CPU and GTX-1080 GPU. As mentioned in Section 3, we choose Resnet-50 model as the backbone of the main branch and we replace the FC layer after the pool5 layer with several $1 * 1$ convolution layers for different classification tasks. We use Softmax layer to process the output of the $1*1$ convolution layer to obtain the final classification results. For the pose branch, we use the MSCOCO dataset to retrain a four-stage model.

In the training process, we divided it into several phases for a better convergence. In the first phase, we train the main branch with the initialized weights pretrained on ImageNet and only the identity classifier is used. The learning rate is initialized to 0.0001 and decayed by 0.96 every epoch. 100 epochs are trained in total. After the first phase, we append the attributes discerning task to the main branch, and train this model in 60 epochs based on the first phase. Finally, we append batch normalization layer after the pool5 layer in main branch and the final stage of pose branch. Then, we use compact bilinear pooling to fuse the pose branch and the trained main branch and then we train this final model 100 epochs. In our experiments, we adopt re-ranking method from [Zhong *et al.*, 2017] to enhance the value of mAP. During training, we set the learning rate to 0.0001 and decay it by 0.96 every epoch. In all these training phases, the Adam optimizer is implemented at the recommended parameters in each mini-batch to update the weights.

4.4 Comparison with State-of-the-art Methods

As our analysis in this paper, the proposed LRDNN can handle human pose change, background redundancy, etc. well. In this section, we compare our LRDNN with the state-of-art methods on Market1501 and DukeMTMC-reid. The methods we adopted to compare include PIE [Zheng *et al.*, 2017a], AttIDNet [Lin *et al.*, 2017], ResNet+OIM [Xiao *et al.*, 2017], ACRN [Schumann and Stiefelhagen, 2017], SVD_ Net [Sun *et al.*, 2017], Part Aligned [Zhao *et al.*, 2017b], PSE [Sarfraz *et al.*, 2018], MGCAM [Song *et al.*, 2018], and AACN [Xu

Methods	Market1501		DukeMTMC-reid	
	Rank-1	mAP	Rank-1	mAP
PIE	79.3	56.0	-	-
AttIDNet	-	-	70.7	51.9
ResNet+OIM	82.1	-	-	68.1
ACRN	83.6	62.6	72.6	52.0
SVD_Net	82.3	62.1	76.7	56.8
Part Aligned	81.1	63.4	-	-
PSE	87.7	69.0	79.8	62.0
MGCAM	83.8	74.3	-	-
AACN	85.9	66.9	76.8	59.3
Ours	**90.4**	**82.8**	**85.3**	**73.2**

Table 2: Quantitative comparison results on Market1501 and DukeMTMC-reid.

et al., 2018]. As shown in Table 2, our method achieves better results in rank-1 value and mAP accuracy. Compared to [Zheng *et al.*, 2017a], the rank-1 value of our method outperforms more than 20% on Market1501 and the mAP accuracy of our method is better than any of the compared methods.

Among the compared methods, [Sarfraz *et al.*, 2018; Lin *et al.*, 2017] adopt attribute clue to guide the deep network to learn attribute recognition which help re-identifying different persons and [Song *et al.*, 2018] use pose clue to learn the local features of human body. Different from these two methods, we aggregate both attribute and pose clue of human body into one framework effectively, and we proposed CPB to enable the deep network to assign different weights to different local parts automatically.

4.5 Ablation Analysis

To further verify the effectiveness of our approach, we compare the contribution of different components in our method by ablation. For the attribute clue and the pose clue, we perform ablation study on both Market1501 and DukeMTMC-reid. The ablation study of the CPB model is performed on Market1501 for a clearer comparison.

Firstly, we investigate the effect of the attribute clue fused into the main branch and the result is listed in Table 3. In this part, we remove the pose branch from the overall frame. As shown in Table 3, the attribute clue (B+A) brings a 6.9% improvement to our algorithm. This is consistent with our intuition that the attribute clue can help person re-ID.

Secondly, we use different kinds of output in the pose branch to find out how pose clue affects the final accuracy. In Table 3, the result of incorporating the part confidence map into the main branch (B+A+H) shows higher accuracy than the part affinity fields (B+A+C), but when incorporating both features, the Rank-1 accuracy achieves 90.4%.

Thirdly, we analyze the influence of the CPB model on Market1501. As shown in Table 3, we observe 0.7% and 1.3% improvement in rank-1 and mAP, respectively, which shows the CPB model has a positive effect on our model.

5 Conclusions

Method	Market1501		DukeMTMC-reid	
	Rank-1	mAP	Rank-1	mAP
B	80.3	69.8	74.1	65.5
B+A	87.2	78.4	82.0	70.6
B+A+H	88.7	80.1	81.4	71.2
B+A+C	87.4	79.6	79.7	70.9
B+A+H+C	**90.4**	**82.8**	**85.3**	**73.2**

Table 3: The ablation analysis results. ”B”, ”A”, ”H” and ”C” means baseline, attribute clue, the part confidence maps from pose branch and the part affinity fields from pose branch.

Market1501	Rank-1	mAP
Ours	**90.4**	**82.8**
Ours without CPB	89.7	81.5

Table 4: The ablation analysis results of the CPB model.

In this paper, we design a local-refining based deep neural network (LRDNN) for person re-ID with attribute discerning. For the complicated human pose and misalignment, we make use of the attribute clue to assist person re-ID in the main branch. Then, we aggregate pose clues generated by a pose branch into the main branch to further excavate local features of human body which reduce the impact of redundant backgrounds on identification tasks. Besides, the CPB model further purifies the pose clue from the pose branch. Finally, the experimental results on the Market1501 and DukeMTMC-reid datasets have verified the effectiveness of our method.

Acknowledgments

This work is supported by Natural Science Foundation of China (No. 61572205 and 61802135), Natural Science Foundation of Fujian Province (No. 2017J01113). The work of Xiangyuan Lan is partially supported by Hong Kong Baptist University Tier 1 Grant.

References

[Cao *et al.*, 2018] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[Felzenszwalb *et al.*, 2010] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[Feng *et al.*, 2018] Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. Learning view-specific deep networks for person re-identification. *IEEE Transactions on Image Processing*, 27(7):3472–3483, 2018.

[Hu *et al.*, 2017] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[Johnson *et al.*, 2018] Jubin Johnson, Shunsuke Yasugi, Yoichi Sugino, Sugiri Pranata, and Shengmei Shen. Person re-identification with fusion of hand-crafted and deep pose-based body region features. *arXiv preprint arXiv:1803.10630*, 2018.

[Kalayeh *et al.*, 2018] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2017] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[Liu *et al.*, 2017] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *arXiv preprint arXiv:1709.09930*, 2017.

[Sarfraz *et al.*, 2018] 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 *Proc. CVPR*, pages 420–429, 2018.

[Schumann and Stiefelhagen, 2017] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1435–1443. IEEE, 2017.

[Song *et al.*, 2018] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.

[Su *et al.*, 2016] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European conference on computer vision*, pages 475–491. Springer, 2016.

[Suh *et al.*, 2018] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. *arXiv preprint arXiv:1804.07094*, 2018.

[Sun *et al.*, 2017] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3800–3808, 2017.

[Sun *et al.*, 2018] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[Wang *et al.*, 2017] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. 2017.

[Xiao *et al.*, 2017] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017.

[Xu *et al.*, 2018] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *arXiv preprint arXiv:1805.03344*, 2018.

[Zhao *et al.*, 2017a] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017.

[Zhao *et al.*, 2017b] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239–3248, 2017.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[Zheng *et al.*, 2016] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[Zheng *et al.*, 2017a] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.

[Zheng *et al.*, 2017b] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 3, 2017.

[Zheng *et al.*, 2018] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[Zhong *et al.*, 2017] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.

[Zhou *et al.*, 2018] Qinqin Zhou, Bineng Zhong, Yulun Zhang, Jun Li, and Yun Fu. Deep alignment network based multi-person tracking with occlusion and motion reasoning. *IEEE Transactions on Multimedia*, 2018.