Black Re-ID: A Head-shoulder Descriptor for the Challenging Problem of Person Re-Identification

Boqiang Xu*
University of Chinese Academy of Sciences
boqiang.xu@cripac.ia.ac.cn

Lingxiao He†
AI Research of JD.com
helingxiao3@jd.com

Xingyu Liao
AI Research of JD.com
liaoxingyu5@jd.com

Wu Liu‡
AI Research of JD.com
liuwu1@jd.com

Zhenan Sun
Institute of Automation, Chinese of Academy of Sciences
znsun@nlpr.ia.ac.cn

Tao Mei
AI Research of JD.com
tmei@jd.com

ABSTRACT

Person re-identification (Re-ID) aims at retrieving an input person image from a set of images captured by multiple cameras. Although recent Re-ID methods have made great success, most of them extract features in terms of the attributes of clothing (e.g., color, texture). However, it is common for people to wear black clothes or be captured by surveillance systems in low light illumination, in which cases the attributes of the clothing are severely missing. We call this problem the *Black Re-ID* problem. To solve this problem, rather than relying on the clothing information, we propose to exploit head-shoulder features to assist person Re-ID. The head-shoulder adaptive attention network (HAA) is proposed to learn the head-shoulder feature and an innovative ensemble method is designed to enhance the generalization of our model. Given the input person image, the ensemble method would focus on the head-shoulder feature by assigning a larger weight if the individual insides the image is in black clothing. Due to the lack of a suitable benchmark dataset for studying the Black Re-ID problem, we also contribute the first Black-reID dataset, which contains 1274 identities in training set. Extensive evaluations on the Black-reID, Market1501 and DukeMTMC-reID datasets show that our model achieves the best result compared with the state-of-the-art Re-ID methods on both Black and conventional Re-ID problems. Furthermore, our method is also proved to be effective in dealing with person Re-ID in similar clothing. Our code and dataset are avaliable on https://github.com/xbq1994/.

*This work is done when Boqiang Xu is an intern at JD AI Research.
†Corresponding author.
‡Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414056

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; **Image representations**; • **Computer systems organization** → **Neural networks**.

KEYWORDS

Black person re-identification, Head-shoulder descriptor, Adaptive attention

ACM Reference Format:
Boqiang Xu, Lingxiao He, Xingyu Liao, Wu Liu, Zhenan Sun, and Tao Mei. 2020. Black Re-ID: A Head-shoulder Descriptor for the Challenging Problem of Person Re-Identification. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414056

1 INTRODUCTION

Person re-identification (Re-ID) is one of the main tasks in computer vision, with the purpose of retrieving the same person from overlapping cameras. In practice, it is common for people to wear black clothes or for clothes to appear black because captured by cameras in low light illumination, in which cases the attributes of the clothing will be lost considerably, posing great challenges to modern person re-identification algorithms. For example, as shown in Figure 1, when people are in black clothes or low light illumination, the seriously loss of clothing attributes causes great difficulties for person retrieval. It is very difficult to distinguish them based only on clothing appearance. We call the process of person Re-ID under black clothing or low light illumination the *Black Re-ID* problem. In addition, We have counted the ids of the people in black clothing in the training set of Market1501 [38], DukeMTMC-reID [22] and CUHK03 [14] in Table 1. From the result we can see that Black Re-ID problem is common in reality.

Recently, methods based on the deep learning networks have made great success in person re-identification by learning powerful person representations [16, 17]. Deep learning based methods typically work on extracting a global feature of a person image [33], which is robust but overlooking the intra-class variations (e.g. occlusions, poses, backgrounds). To alleviate this problem, the local features are leveraged to obtain a more discriminative representation for effective person Re-ID. Several pose-based methods [35, 37]

Figure 1: The illustration of the Black Re-ID problem. Black Re-ID poses a huge challenge to person re-identification due to the little information about clothes. Head-shoulder feature obtains abundant discriminative information for contributing the person Re-ID, such as gender, haircut, appearance and glasses.

Table 1: The ids of the people in black clothing in the training set of Market1501 [38], DukeMTMC-reID [22] and CUHK03 [14]

	Market1501	DukeMTMC-reID	CUHK03
# identities_total	751	702	767
# identities_black	98	347	331

attend to extract local features from body parts such as arms, legs and torso. They utilize off-the-shelf pose estimators to enhance the attention to the corresponding local parts. Some part-based models [29, 31] extract local detail information by slicing the feature into several horizontal grids, training them individually and aggregating them for a powerful representation. However, these methods extract features mainly rely on the attributes of the clothing (e.g. color, texture, style). In Black Re-ID problem, the information lost due to the black clothing or insufficient illumination hinders the performance of these Re-ID methods. As shown in Figure 1, it is very difficult to distinguish them based only on the clothes.

In this paper, we propose to leverage the head-shoulder information to enhance the description of the feature for people in black clothing for an more effective person Re-ID. Head-shoulder part possesses abundant discriminative information, such as haircut, face and other appearances, for contributing the person Re-ID in solving the Black Re-ID problem. In particular, as shown in Figure 2, we design a two-stream network consists of the global stream and head-shoulder attention stream (HSA). For the global stream, it learns the global representation. For the head-shoulder attention stream (HSA), it learns the head-shoulder feature and utilizes the head-shoulder localization layer to take place of the off-the-shelf pose estimators with lightweight architecture. Furthermore, in order to make use of the head-shoulder information in solving both Black Re-ID and conventional problems, we propose the adaptive attention module to adapt the weights of the global and head-shoulder feature by the condition of the input person image. Specially, when the input person is in black clothing, our model would give more attention to the head-shoulder feature, compared to the person not in black clothes. This assembled method ensures that our model can make the most efficient use of the head-shoulder information in both the Black and conventional Re-ID conditions. In addition, we establish the first Black-reID dataset which contains 1274 identities in training set. Our experiment was conducted on the Black-reID dataset, Market1501 [38] and DukeMTMC-reID [22], demonstrating the advantage of our approach for person Re-ID on both Black and conventional Re-ID problems.

The main contributions of this paper can be summarized as follows:

- We firstly propose the study of the Black Re-ID problem and establish the first Black-reID dataset.
- We propose the head-shoulder adaptive attention network (HAA), which makes use of the head-shoulder information to support person re-identification through the adaptive attention module. The HAA can be integrated with the most current Re-ID framework and is end-to-end trainable.
- Our model is proved by the experiment that it is not only effective for Black Re-ID problem but also valid in similar clothing.
- We achieve a new state of the art, and outperforms other Re-ID methods by a large margin in solving both Black Re-ID and conventional problems.

2 RELATED WORK

Person Re-ID. The performance of person Re-ID has made great success recently [5, 6, 13, 23], due to the development of **CNN**s. Most of the **CNN**s based methods treat Re-ID as a classification task [41], aiming at divide person with same identity into the same category. To obtain a more discriminative and robust representation of the person, many methods [29, 31, 35–37] integrate global feature with local features for an effective person re-identification.

Local feature based methods enhance the discriminative capability of the final feature map. We briefly classify them into three categories: The first approach is pose-based Re-ID [35, 37], which uses an off-the-shelf pose estimator to extract the pose information for aligning body parts [28] or generating person images [15]. However, the pose estimator is utilized through the whole training and testing process, making the network bigger and slower. The second approach is part-based Re-ID [29, 31], which slices image or global feature into several horizontal grids, training individually, and assembling for a discriminative person representation. However, this method is sensitive to the pose variations and occlusions, as they may compare corresponding part with different semantics. The third kind of methods leverage local information with attention maps [36], which can be trained under less supervisory signals compared to the pose annotations. It pays more attention to the regions of interest and robust to the background clutter, but nonetheless, the area of concern may not contain body parts. Our method belongs to the first category. In contrast to the other pose-based methods, we propose the *HSA* stream to take place of the pose estimator, making the network lightweight. Furthermore, we introduce an adaptive attention module to decide the weight of global and local features by the condition of the input person image.

Head-shoulder Information. Head-shoulder information is reliable for retrieving as the corresponding region can usually be captured in reality, while the whole body may be occluded. Moreover, the head-shoulder feature possesses abundant discriminative information for contributing the person Re-ID, such as haircut, complexion or appearance. Unfortunately, as we know, there are few works [12] focusing on improving Re-ID performance with head-shoulder part. Li et al. [12] propose a multi pose learning based model which takes head-shoulder part as input, aiming at solving partial Re-ID problem [40] in crowded conditions. It tackles head-shoulder pose variations by learning an ensemble verification conditional probability distribution about relationship among multiple poses. However, it focuses mainly on pose variation, contributing little work to the head-shoulder localization, feature extraction and feature fusion with the global representation. Moreover, it performs worse than other methods (i.e., MGN [31], PCB [29], AlignedReID [18]) when the whole body is available as it only takes head-shoulder part as input.

3 OUR APPROACH

We show in Figure 2 (a) an overview of our framework. As our method could be integrated with the most current Re-ID model (*e.g.*, MGN [31]), the backbone can be selected according to different requirements of accuracy and speed. To illustrate our framework concisely, the ResNet-50 [4] trained for ImageNet classification [2] is exploited in demonstration. The network consists of two streams named global stream and head-shoulder attention stream (HSA). The first stream extracts the global feature from the person image. The second stream focuses on localizing and extracting head-shoulder information to make the final feature more discriminative for dealing with the Black Re-ID problem. Further more, we propose the adaptive attention module to adapt the weights of the global feature and head-shoulder feature, varying depending on whether a person is in black or not. We train our model end-to-end using cross-entropy, triplet and L2 losses. At test time, we extract features from person, and calculate the Euclidean distance between them to match people with the same ID.

3.1 Global Stream

We extract global feature through a ResNet50 fasion network. Specifically, a feature map of size $C \times H \times W$ is extracted from a person image, where C, H, W represent the number of channels, height and width, respectively. The resulting feature map is then processed by GAP and channel reduction to produce the global feature, denoted by f_g, of size $c \times 1 \times 1$,

3.2 Head-shoulder Attention Stream

Head-shoulder Feature Construction. As mentioned in Section 2, several part-based Re-ID models attempt to localize body parts (e.g. legs, arms) with some off-the-shelf pose estimators [35, 37] through both the training and testing phase. However, these methods make the Re-ID model bigger and slower, as it is a combination of two models. For the head-shoulder area, it's a less strictly defined area, and slight offset of localization would not affect Re-ID performance, which gives it the advantage of lower precision requirement of positioning than the segmentation task does. With these concerns, what we need is a lightweight localization layer that can learn a bounding box, representing the head-shoulder region.

We propose the HSA stream to localize and extract the feature of the head-shoulder region, avoiding the use of pose estimators. HSA stream contains a head-shoulder localization layer (HLL), which is inspired by the success of $Spatial\ Transformer\ Network$ (**STN**) [8]. As illustrated in Figure 2 (b), HLL is an end-to-end trainable module with capacity of applying an affine transformation to the feature map, including scaling, translation and rotation. As the head-shoulder region is a simple bounding box, we only retain the ability to zoom and shift, equivalent to the cropping from the input. The transformation of the head-shoulder localization layer can be realized as follows:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

where x_i^s,y_i^s and x_i^t,y_i^t are the source positions and target coordinates of the i-th pixel respectively, s_x,s_y are scaling parameters and t_x,t_y are translation parameters. In the head-shoulder localization layer, the fully connected layer of size $C \times 4$ outputs s_x,s_y,t_x and t_y, and then the following steps will sample pixels from the corresponding positions of the input person image to generate the bounding box.

Specifically, the input person images firstly pass through the head-shoulder localization layer, giving the head-shoulder region and resize to the same shape as input. Then, a feature map of size $C{\times}H{\times}W$ is extracted from the head-shoulder region and sliced into 3 horizontal grids. The head-shoulder attention network(HAN) is applied to each individual horizontal slice, which are concatenated finally to produce the head-shoulder feature f_h of size $c \times 1 \times 1$.

Head-shoulder Attention Network. Figure 2 (c) shows the detailed structure of the HAN. Since different channels of feature maps represent different meanings, that is, the contributions of features to recognition vary from channel to channel, and different spatial location of features has diverse semantics. We introduce an attention network to enhance the representation of head-shoulder in both channel and spatial dimensions.

Specifically, for the i-th($i = 1, 2, 3$) slice, the input feature X_i passes through a gating mechanism, including a generalized mean pooling (GeM) [3], a fully connected layer with weight $W_i \in \mathbb{R}^{C \times \frac{C}{r}}$ for dimension reduction, a ReLU activation, another fully connected layer with weight $U_i \in \mathbb{R}^{\frac{C}{r} \times C}$ for dimension incrementation and a sigmoid activation σ. Here, r is the reduction ratio. Then, the channels are reweighted by a shortcut connection architecture with element-wise addition, which can be formulated as:

$$A_i = X_i + X_i \bullet d_i \tag{2}$$

$$d_i = \sigma(U_i RELU(W_i X_i)) \tag{3}$$

where $| \bullet |$ is element-wise multiplication and A_i is the output after channel attention. For the description, A_i can be written as:

$$[A_i^1, A_i^2, A_i^3, ...A_i^C] \tag{4}$$

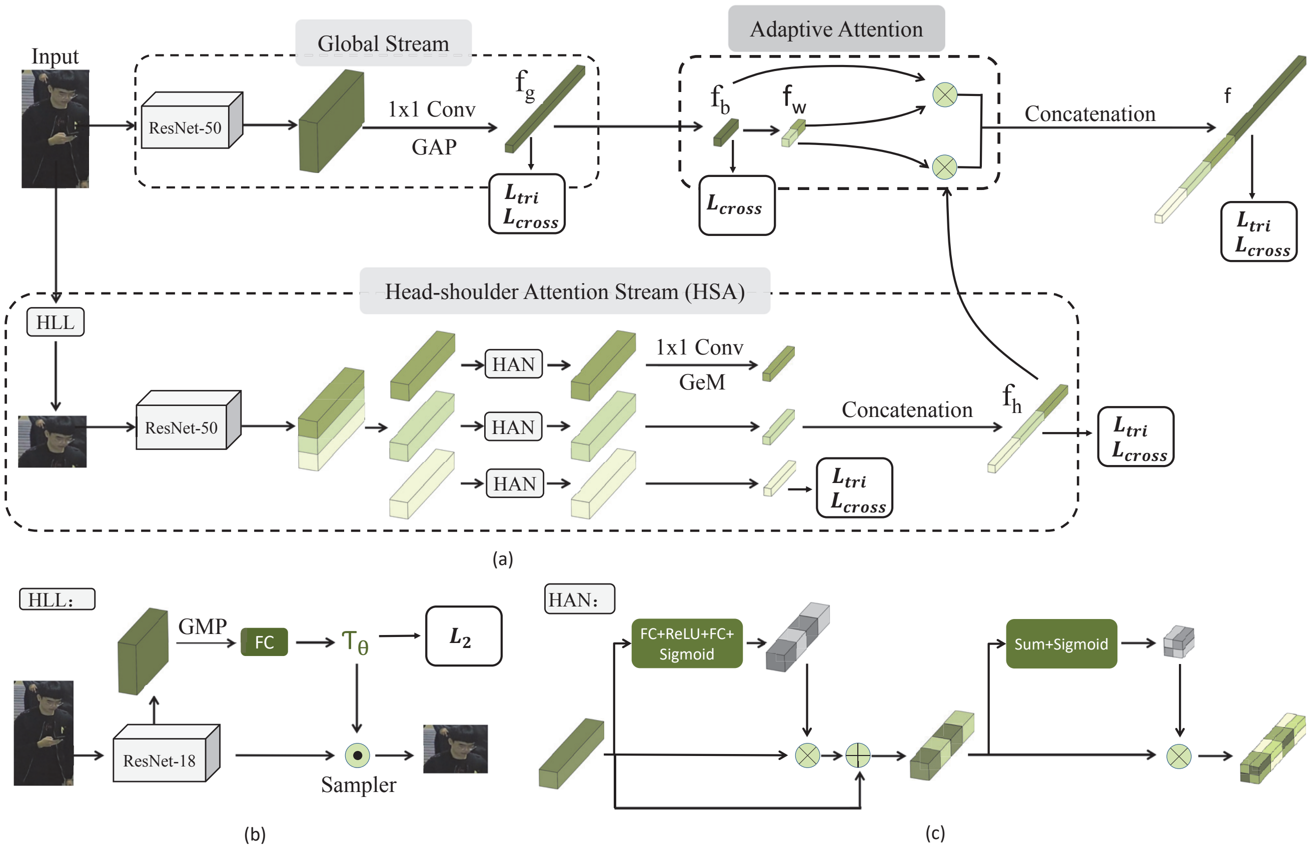

Figure 2: Overview of the proposed framework. (a) Our model consists of two streams: the global stream extracts the global feature from the input person image; the head-shoulder attention stream (HSA) crops the head-shoulder region by head-shoulder localization layer (HLL), which is divided into three horizontal stripes later and fed into head-shoulder attention network (HAN) to give the head-shoulder representation; at the end of the model, global and head-shoulder features are assembled through the adaptive attention module to produce the final representation for Re-ID. (b) The detailed structure of the head-shoulder localization layer (HLL). (c) The detailed structure of the head-shoulder attention network (HAN). Here, $\otimes$, $\oplus$, GAP, GMP, GeM, $L_{triplet}$, L_{ce}, L_2 indicate element-wise multiplication, element-wise addition, global average pooling, global max pooling, generalized mean pooling, triplet loss, cross-entropy loss and L2 loss respectively, and $\odot$ is a sampler.

where $A_i^n (n = 1, 2, ...C)$ are the features of each channel of A_i. The spatial attention is conducted by strengthening the peak responses, this process can be formulated as:

$$f_{hi} = A_i \bullet \xi(\sum_{n=1}^{C} A_i^n) \tag{5}$$

3.3 Adaptive Attention

Most of the existing Re-ID methods [29, 31] directly concatenate global and local features, ignoring the relationship between the weight of the features and input conditions. That is, no matter what kind of person is input such as occluded or exposed, the network gives the same attention to the global feature as to the local feature. To alleviate this problem, we propose the adaptive attention module to determinate the global and local feature weights by distinguishing input types. Specifically, the adaptive attention stream would decide if it is a person in black first, and give more attention to head-shoulder feats for people in black clothes than that not in black.

Concretely, firstly, the global feature f_g is fed into a fully connected layer to gather the feature of size $N \times 2$, where N is the batch size, denoted by f_b, representing whether the input person is in black. After that, f_b is fed into another fully connected layer, giving feature map f_w of size $N \times 2$. f_w is the weights of global feature and head-shoulder feature, varying depending on whether the person is in black or not. That is, higher attention would be applied to the head-shoulder feature when the person is in black clothes. Finally, we integrate the global feature and head-shoulder feature as follows:

$$f = (f_g \bullet w_1) \circledast (f_h \bullet w_2) \tag{6}$$

where $f_w = \begin{bmatrix} w_1 & w_2 \end{bmatrix}$, $| \bullet |$ is element-wise multiplication, f_g and f_h are global feature and head-shoulder feature respectively and

$\circledast$ means concatenate method. The feature f is used as the person representation for Re-ID.

3.4 Model Training

To train our model, we use triplet and cross-entropy loss, balanced by the parameter α and β as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{triplet} + \beta\mathcal{L}_{ce}$$

where we denote by $\mathcal{L}_{triplet}$ and $\mathcal{L}_{ce}$, triplet and cross-entropy losses, respectively. The cross-entropy loss is defined as

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N} \log \frac{\exp(\mathbf{W}_{y_i}\mathbf{h}_p^i + b_{y_i})}{\sum_{k=1}^{C}\exp(\mathbf{W}_k\mathbf{h}_p^i + b_k)} \tag{8}$$

where N is the number of images in mini-batch, y_i is the label of feature h_p^i and C is the number of classes. Given a triplet consisting of the anchor, positive, and negative features (i.e., $\mathbf{q}_{i,j}^A$, $\mathbf{q}_{i,j}^P$ and $\mathbf{q}_{i,j}^N$), the batch-hard triplet loss [7] is formulated as follows:

$$\mathcal{L}_{triplet} = \sum_{k=1}^{N_k}\sum_{m=1}^{N_M}[\alpha + \max_{n=1\ldots M}||\mathbf{q}_{k,m}^A - \mathbf{q}_{k,n}^P||_2 - \min_{\substack{l=1\ldots K\\ n=1\ldots N\\ l\neq k}}||\mathbf{q}_{k,m}^A - \mathbf{q}_{l,n}^N||_2]_+ \tag{9}$$

where α denotes the margin. To supervise the study of the head-shoulder localization layer and minimize the regression error of it, L2-loss is adopted and defined as follows:

$$\mathcal{L}_2 = \frac{1}{2N}\sum_i ||l^{(i)} - r(h^i)||_2^2 \tag{10}$$

where N is the batch size, $l^{(i)}$ and $h^{(i)}$ are the ground-truth label and the prediction of the i-th bounding box of the head-shoulder region and r is the function that transform $h^{(i)}$ to the same coordinate system as $l^{(i)}$.

4 THE BLACK-REID DATASET

To promote the research on the Black Re-ID problem, we introduce Black-reID, the first dataset focusing on Black Re-ID problem, derived from the Market1501 [38], DukeMTMC-reID [22], Partial [39] and Occluded [44] datasets.

4.1 Properties of Black-reID Dataset

Some examples of the Black-reID dataset are shown in Figure 3. There are a few advantages to make Black-reID appealing. First, it is the first dataset for Black Re-ID problem. According to the clothes that the people wear, the Black-reID dataset consists of two groups. The first group is black group, which focuses on the Black-reID problem and contains 5,649 images covering 688 identities in training set and 1,644 identities of 1,489 query images and 4,973 gallery images in test set. The second group is white group, which is constructed for investigating the Re-ID ability of the model in similar clothing. The white group contains 10,040 images of 586 subjects for training, 2,756 images of 628 subjects for query and 10336 images for gallery. Second, to fit the reality, both two groups include people wearing the corresponding color and those who are not. The reason for this design is that we want our method can not only solve the Black Re-ID problem effectively, but also be reliable when dealing with conventional scenarios. Third, we have labeled the bounding box of head-shoulder region for the training set in Black-reID and annotated ids of person in black and white clothes for the two groups, respectively.

4.2 Data Collection

We pick out 98 ids, 347 ids, 5 ids and 42 ids who are in black clothes from Market1501 [38], DukeMTMC-reID [22], Partial [39] and Occluded [44] datasets respectively for the black group, and 665 pedestrians who are in white from Market1501 [38] for the white group. Moreover, we randomly add some people in other clothes to these two groups for creating the Black-reID dataset.

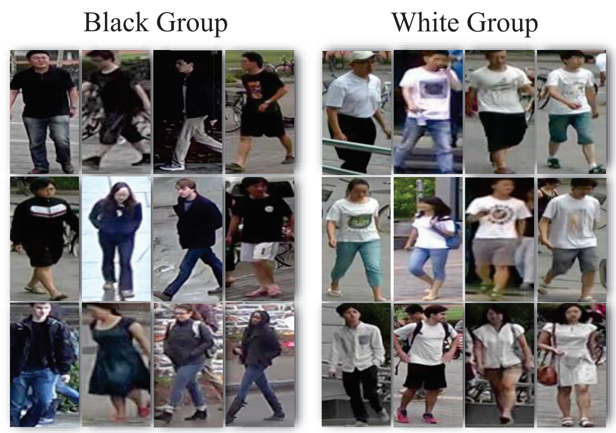

Figure 3: Some examples of the Black-reID dataset

5 EXPERIMENT

5.1 Implementation details

Dataset Our method is tested on the following three datasets and compare with the state-of-the-art methods. 1) The Black-reID dataset established by us, which contains 1274 identities in training set. 2) The Market1501 dataset [38] consists of 32,688 images of 1,501 subjects captured by six cameras. 3) The DukeMTMC-reID [22] offers 16,522 training images of 702 identities, 2,228 query and 17,661 gallery images of 702 identities. Additionally, we manually labeled the head-shoulder bounding box for the Market1501 and DukeMTMC-reID datasets.

Training. We resize all the training images into 384×128. We set the number of feature channels c to 1536 and batch size N to 64. We adopt horizontal flipping and random erasing [43] for data augmentation. Following the work of [29], the GAP and fully connected layers are removed from the original ResNet-50 architecture and the stride of the last convolutional layer is set to 1. Firstly, we pretrain the head-shoulder localization layer on Black-reID dataset and freeze it in the following training. Next, we train the global stream and head-shoulder attention stream for 50 epochs individually, and then train them together with adaptive attention module for another 40 epochs. We use the adaptive gradient (**Adam**) [10] as the optimizer with β_1 of 0.9, β_1 of 0.999 and weight decay of

5e-4. The learning rate is initially set to 3e-4 and divided by 10 at 40 and 70 epochs. We set the weight parameter α and β to 1,1 and 1,2 for the integration of our model with ResNet50 [4] and MGN [31], respectively.

Evaluation Metrics. We utilize mean average precision (mAP) and Cumulative Matching Characteristic (CMC) curves to evaluate the performance of various Re-ID models. All the experiments are conducted in a single query setting.

Table 2: Quantitative comparison with the state-of-the-art methods in person re-id on Black-reID dataset. Bold number denote the best performance. We denote HAA (ResNet50) and HAA (MGN) by the method selecting ResNet50 and MGN as the backbone respectively.

Method	Black Group		White Group	
	mAP	Rank-1	mAP	Rank-1
ResNet50 [4]	70.8	80.9	75.8	89.5
PCB [29]	73.4	83.2	78.2	90.8
AlignedReID [34]	75.5	83.5	80.5	91.3
MGN [31]	79.1	86.7	85.8	94.3
HAA (ResNet50)	79.0	86.7	84.4	93.5
HAA (MGN)	**83.8**	**91.0**	**88.1**	**95.3**

5.2 Comparison with the state of the art

Results on Black-reID. We compare in Table 2 our method with the state-of-the-art Re-ID models (i.e., MGN [31], PCB [29], AlignedReID [18]) on Black-reID dataset. We denote HAA (ResNet50) and HAA (MGN) by the method selecting ResNet50 and MGN as the backbone respectively. Table 2 shows that HAA (ResNet50) and HAA (MGN) outperform their control groups (i.e., ResNet50 and MGN, respectively) by a large margin. In the black group, HAA (MGN) achieves the best result with mAP of 83.8% and rank-1 accuracy of 91%, which is 4.7% and 4.3% higher than the corresponding metrics of MGN respectively. HAA (ResNet50) also gives 8.2% and 5.8% higher points in mAP and rank-1 than ResNet50, which reaches the same level as MGN. This demonstrates the effectiveness of our HAA in dealing with the Black Re-ID problem. In the white group, we can see that HAA (ResNet50) gives 8.6% and 4% higher points in mAP and rank-1 accuracy than ResNet50 while HAA (MGN) achieves the best result with mAP of 88.1% and rank-1 accuracy of 95.3%, which is 2.3% and 1% higher than the corresponding metrics of MGN respectively. The result proves that our model is not only effective for Black Re-ID problem but also valid in solving similar clothes. We compare some retrieved results between PCB [29], AlignedReID [18], MGN [31] and our method in Figure 4. As shown in Figure 4, PCB [29], AlignedReID [18] and MGN [31] extracts features mainly rely on the clothing attributes. By contrast, we can see that our proposed HAA can improve the Re-ID performance on Black Re-ID problem by taking advantage of the head-shoulder attributes to make the representation more discriminative. Furthermore, the Re-ID performance in Black -reID dataset is lower than that in Market1501 and DukeMTMC-reID. This proves that when people are in black clothes, lacking of attributes of clothing, the Re-ID performance of the model is really degraded.

Results on Market1501 and DukeMTMC-reID. We also compare our proposed method with current state-of-the-art methods of three categories on Market1501 and DukeMTMC-reID in Table 3. *Pose − guided methods* leverage pose information to extract more discriminative local details or align features. *Part − based methods* slice the image/feature maps into several horizontal grids to assist Re-ID. *Attention − based methods* compute attention maps to consider discriminative regions of interest. Note that we do not utilize re-ranking [42] in all our results for clear comparisons.

From Table 3, we can see that HAA (Ours) gives the best results on all datasets and specifically gives big improvement to mAP. HAA (Ours) achieves mAP of 89.5% and rank-1 accuracy of 95.8% on the Market1501, which is 2.6% and 0.1% higher than the corresponding metrics of MGN respectively, and mAP of 80.4% and rank-1 accuracy of 89% for the DukeMTMC-reID, which is 2% and 0.3% higher than the MGN respectively.

6 ABLATION STUDY

We conduct extensive ablation studies on the black group of the Black-reID dataset based on the HAA (ResNet50).

The Impact of the Adaptive Attention Module. In our method, we concatenate the global and head-shoulder feature through the adaptive attention module. To show how the ensemble method works, we set up a control group where the global and head-shoulder feature are directly concatenate during the whole training and testing process. The result is illustrated in Table 4. We denote HAA (ResNet50) (w/o adaptive attention) and HAA (ResNet50) (adaptive attention) by directly concatenate the global and head-shoulder feature and concatenate them through adaptive attention module, respectively. From the results, we can see that the adaptive attention stream gives the performance gains of 0.7% and 0.5% for mAP and rank-1 accuracy on Black-reID, 0.3% and 0.4% for mAP and rank-1 accuracy on Market1501 and 1.1% and 0.8% for mAP and rank-1 accuracy on DukeMTMC-reID,respectively. The result proves the effectiveness of the adaptive attention module in solving both Black and conventional Re-ID problems.

Ablation Study of the GeM Pooling. We adopt GeM pooling in the HSA stream. To compare the effectiveness of diverse pooling methods, we replace the GeM pooling with GAP and GMP, and conduct the experiments on Black-reID dataset in Table 5. From the third and fourth rows, we can see that GMP gives better results than GAP, the reason is that GAP covers the whole person image and is easily distracted by background and black-clad while GMP gets over this problem by aggregating the feature from the most discriminative part. The results in the next row show that GeM pooling achieve the best result, giving the performance gains of 1.7%, 1.4% and 3.8%, 4% for mAP and rank-1 accuracy than GMP and GAP respectively. GeM pooling is given by:

$$e = [(\sum_{u\in\sigma} x_{cu}^{p})^{\frac{1}{p}}]_{c=1..C} \tag{11}$$

Table 3: Performance (%) comparisons with the state-of-the-art methods on Market1501 and DukeMTMC-reID.

	Method	Market1501		DukeMTMC-reID	
		mAP	Rank-1	mAP	Rank-1
Basic-CNN	ResNet50 [4]	84.6	93.3	75.3	86.2
Pose-guided methods	Spindle [35]	-	76.9	-	-
	PIE [37]	53.9	78.7	-	-
	MSCAN [11]	57.5	80.8	-	-
	PDC [27]	63.4	84.1	-	-
	Pose Transfer [15]	68.9	87.7	48.1	68.6
	PN-GAN [21]	72.6	89.4	53.2	73.6
	PSE [24]	69.0	87.7	62.0	79.8
	MGCAM [26]	74.3	83.8	-	-
	MaskReID [19]	75.3	90.0	61.9	78.9
	Part-Aligned [28]	79.6	91.7	84.4	69.3
	AACN [32]	66.9	85.9	59.3	76.8
	SPReID [9]	81.3	92.5	71.0	84.4
Part-based methods	AlignedReID [34]	79.3	91.8	-	-
	Deep-Person [1]	79.6	92.3	64.8	80.9
	PCB+RPP [29]	81.6	93.8	69.2	83.3
	MGN [31]	86.9	95.7	78.4	88.7
Attention-based methods	DLPAP [36]	63.4	81.0	-	-
	HA-CNN [20]	75.7	91.2	63.8	80.5
	DuATM [25]	76.6	91.4	64.6	81.8
	Mancs [30]	82.3	93.1	71.8	84.9
	HAA (Ours)	**89.5**	**95.8**	**80.4**	**89.0**

Table 4: Ablation study of the adaptive attention module on Black-reID, Market1501 and DukeMTMC-reID datasets. w/o: without.

Methods	Black-reID		Market1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
HAA (ResNet50) (w/o adaptive attention)	78.3	86.2	85.3	93.8	73.1	85.6
HAA (ResNet50) (adaptive attention)	**79.0**	**86.7**	**85.6**	**94.2**	**74.2**	**86.4**

Table 5: Ablation study of the pooling methods on Black-reID dataset. GAP, GMP, GeM indicate global average pooling, global max pooling and generalized mean pooling, respectively.

	mAP	Rank-1
ResNet50 [4]	70.8	80.9
HAA (ResNet50) (GAP)	75.2	82.7
HAA (ResNet50) (GMP)	77.3	85.3
HAA (ResNet50) (GeM Pooling)	**79.0**	**86.7**

where $u \in \sigma = 1, ..., H \times 1, ..., W$ is a 'pixel' in the map, c is channel, x_{cu} is the corresponding tensor element and p is a learnable parameter. GeM pooling is a generation of the GAP($p = 1$) and GMP($p = \infty$) and the larger the p the more localized the feature map responses are. GeM pooling can learn the p to aggregate the feature from the appropriate size area, which is between the 'whole

Table 6: Quantitative comparison of diverse features on the Black-reID dataset

	mAP	Rank-1
ResNet50 [4]	70.8	80.9
HAA (ResNet50) (Global)	78.6	86.4
HAA (ResNet50) (Head-shoulder)	44.8	51.5
HAA (ResNet50) (Global+Head-shoulder)	**79.0**	**86.7**

Figure 4: Comparison of the PCB [29], AlignedReID [18], MGN [31] and our proposed HAA in solving Black Re-ID problem. Blue and red rectangles indicate correct and error retrieval results,respectively. More blue/correct retrievals in HAA result show that the head-shoulder information is discriminative for better Re-ID results.

image' and 'pixel', of the feature map. Therefore, we adopt GeM pooling in the HSA stream.

Performance comparison of global and head-shoulder features. We demonstrate the capabilities of different features for person Re-ID in Table 6. We retrieve person images from Black-reID dataset using global feature, head-shoulder feature and integration of them in the experiment, which are denoted by HAA (ResNet50) (Global), HAA (ResNet50) (Head-shoulder) and HAA (ResNet50) (Global+Head-shoulder), respectively. From this table, we can observe two things: (1) The integration of global and head-shoulder feature achieves the best result in person Re-ID. HAA (ResNet50) (Global+Head-shoulder) gives 0.4%, 0.3% and 34.2%, 35.2% higher points in mAP and rank-1 accuracy than HAA (ResNet50) (Global) and HAA (ResNet50) (Head-shoulder), respectively. (2) Our model improves the representative capability of the global feature drastically. The improvements in mAP and rank-1 accuracy are 7.8% and 5.5% respectively.

7 CONCLUSION

We have presented a head-shoulder adaptive attention network to support person Re-ID with head-shoulder information. Through the adaptive attention module, the weights of the global and head-shoulder features can be automatically adjusted based on the type of the input person image. Our head-shoulder adaptive attention network can be integrated with the most current Re-ID models and is end-to-end trainable. We have also firstly proposed the Black Re-ID challenge and the first Black-reID dataset for further research. Our method not only achieves the best performance on Black-reID dataset but also on Market-1501 and DukeMTMC-reID and is proved to be effective in dealing with similar clothing. On Black-reID dataset, our model significantly outperforms the previous methods, by at least **+4.7%/4.3%** in mAP/rank-1 accuracy.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. U1836217). The code was developed based on the 'fast-reid' toolbox https://github.com/JDAI-CV/fast-reid.

REFERENCES

[1] Xiang Bai, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu. 2020. Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. *Pattern Recognit.* 98 (2020).

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), 248–255.

[3] Yinzheng Gu, Chuanpeng Li, and Jinbin Xie. 2018. Attention-aware Generalized Mean Pooling for Image Retrieval. *ArXiv* abs/1811.00202 (2018).

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 770–778.

[5] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7073–7082.

[6] Lingxiao He, Yinggang Wang, Wu Liu, Xingyu Liao, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware Pyramid Reconstruction for Alignment-free Occluded Person Re-identification. *Computer Vision and Pattern Recognition (CVPR)* (2019).

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *ArXiv* abs/1703.07737 (2017).

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. In *NIPS*.

[9] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human Semantic Parsing for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 1062–1071.

[10] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2015).

[11] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. 2017. Learning Deep Context-Aware Features over Body and Latent Parts for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 7398–7407.

[12] Jia Li, Yunpeng Zhai, Yaowei Wang, Yemin Shi, and Yonghong Tian. 2018. Multi-Pose Learning based Head-Shoulder Re-identification. *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (2018), 238–243.

[13] Shuangqun Li, Xinchen Liu, Wu Liu, Huadong Ma, and Haitao Zhang. 2016. A discriminative null space based deep learning approach for person re-identification. In *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*.

[14] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 152–159.

[15] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. 2018. Pose Transferrable Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 4099–4108.

[16] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7202–7211.

[17] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-Scale Triplet CNN for Person Re-Identification. In *Proceedings of the 24th ACM International Conference on Multimedia*. 192–196.

[18] Hao Luo, W. W. Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. 2019. AlignedReID++: Dynamically matching local information for person re-identification. *Pattern Recognit.* 94 (2019), 53–61.

[19] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. 2018. MaskReID: A Mask Based Deep Ranking Neural Network for Person Re-identification. *ArXiv* abs/1804.03864 (2018).

[20] Wei qi Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious Attention Network for Person Re-identification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 2285–2294.

[21] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-Normalized Image Generation for Person Re-identification. In *European Conference on Computer Vision (ECCV)*.

[22] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In *European Conference on Computer Vision (ECCV) Workshops*.

[23] Weijian Ruan, Wu Liu, Qian Bao, Jun Chen, Yuhao Cheng, and Tao Mei. 2019. POINet: Pose-Guided Ovonic Insight Network for Multi-Person Pose Tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*. 284–292.

[24] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A Pose-Sensitive Embedding for Person Re-identification with Expanded Cross Neighborhood Re-ranking. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 420–429.

[25] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex Chichung Kot, and Gang Wang. 2018. Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 5363–5372.

[26] Chunfeng Song, Ying Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-Guided Contrastive Attention Model for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 1179–1188.

[27] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2017. Pose-Driven Deep Convolutional Model for Person Re-identification. *IEEE International Conference on Computer Vision (ICCV)* (2017), 3980–3989.

[28] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-Aligned Bilinear Representations for Person Re-identification. *ArXiv* abs/1804.07094 (2018).

[29] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling. In *European Conference on Computer Vision (ECCV)*.

[30] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2018. Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification. In *European Conference on Computer Vision (ECCV)*.

[31] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. *Proceedings of the 26th ACM international conference on Multimedia* (2018).

[32] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-Aware Compositional Network for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 2119–2128.

[33] Li Zhang, Tao Xiang, and Shaogang Gong. 2016. Learning a Discriminative Null Space for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1239–1248.

[34] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. 2017. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. *ArXiv* abs/1711.08184 (2017).

[35] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 907–915.

[36] Liming Zhao, Xiao jie Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-Learned Part-Aligned Representations for Person Re-identification. *IEEE International Conference on Computer Vision (ICCV)* (2017), 3239–3248.

[37] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. 2019. Pose-Invariant Embedding for Deep Person Re-Identification. *IEEE Transactions on Image Processing* 28 (2019), 4500–4509.

[38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. *IEEE International Conference on Computer Vision (ICCV)* (2015), 1116–1124.

[39] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jian-Huang Lai, and Shaogang Gong. 2015. Partial Person Re-Identification. In *IEEE International Conference on Computer Vision (ICCV)*. 4678–4686.

[40] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jian-Huang Lai, and Shaogang Gong. 2015. Partial Person Re-Identification. *IEEE International Conference on Computer Vision (ICCV)* (2015), 4678–4686.

[41] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14 (2017), 1 – 20.

[42] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking Person Re-identification with k-Reciprocal Encoding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 3652–3661.

[43] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2017. Random Erasing Data Augmentation. *ArXiv* abs/1708.04896 (2017).

[44] Jiaxuan Zhuo, Zeyu Chen, Jian-Huang Lai, and Guangcong Wang. 2018. Occluded Person Re-Identification. *IEEE International Conference on Multimedia and Expo (ICME)* (2018), 1–6.