# Person Re-Identification With Reinforced Attribute Attention Selection

Jianfu Zhang, Li Niu, and Liqing Zhang, *Member, IEEE*

***Abstract*— Person re-identification (Re-ID) aims to match pedestrian images across various scenes in video surveillance. There are a few works using attribute information to boost Re-ID performance. Specifically, those methods leverage attribute information to boost Re-ID performance by introducing auxiliary tasks like verifying the image level attribute information of two pedestrian images or recognizing identity level attributes. Identity level attribute annotations cost less manpower and are well-fitted for person re-identification task compared with image-level attribute annotations. However, the identity attribute information may be very noisy due to incorrect attribute annotation or lack of discriminativeness to distinguish different persons, which is probably unhelpful for the Re-ID task. In this paper, we propose a novel Attribute Attentional Block (AAB), which can be integrated into any backbone network or framework. Our AAB adopts reinforcement learning to drop noisy attributes based on our designed reward and then utilizes aggregated attribute attention of the remaining attributes to facilitate the Re-ID task. Experimental results demonstrate that our proposed method achieves state-of-the-art results on three benchmark datasets.**

***Index Terms*— Person re-identification, attentional module, sequential decision making.**

## I. Introduction

GIVEN a query image, the goal of person re-identification (Re-ID) is to identify the images containing the same person from the gallery images captured by non-overlapping surveillance cameras [1]. The previous re-id works either learn discriminative features which are invariant across different camera views and illumination conditions [2]–[5], or learn the distance metric which can preserve the ranking order of training samples [6]–[9]. In recent years, rising deep learning techniques [10]–[12] have been used for the Re-ID task and achieved excellent performance by learning discriminative features and good metric simultaneously in an end-to-end system.

Despite the rapid progress in Re-ID, Re-ID models suffer from misalignment issue due to the pose variation, wrong bounding boxes, and the variations of the camera views. Recently, several methods [13]–[16] have employed attentional models, which extract features from regions of interest, to alleviate misalignment issue and supremely improve the performance. Nevertheless, these methods only learn global attention on the entire human body and cannot capture fine-grained attention related to attributes (*e.g.*, cloth colors, age, gender) while such attribute information has been proved useful for the Re-ID task in recent works [17], [18]. Inspired by fully attentional block proposed in [14], we propose a novel Attribute Attentional Block (AAB), which combines both global and attribute attention to generate attention map.

The attributes mentioned above are ID-level attributes because we assume that all the images belonging to the same ID should have the same attribute value. In Figure 1(a), we represents some examples of ID-level annotations for pedestrian images. In contrast, image-level attributes may be independent of persons, indicating that different images at different surveillance videos from the same person could have different attribute values. Thus, such attributes are unsuitable for Re-ID purpose. Also, image-level attribute annotations cost much more manpower compared with ID-level attribute annotations.

However, there exist several major issues when utilizing attribute information. First, Re-ID images are often of low quality (*e.g.*, low resolution, distracting persons, and wrong bounding boxes), which is likely to induce annotation errors of human annotators. Second, provided attribute annotation is usually ID-level attributes instead of image-level attributes because collecting image-level attributes is very time-consuming and expensive. For different images from the same ID, some attributes may vary along with the time (*e.g.*, vanishing attributes due to the occlusion caused by camera view change), resulting in the inconsistency of ID-level attribute on the image level. Third, some of the attributes are rare (*e.g.*, a few persons wear hats in a cloudy environment) or not discriminative enough to distinguish different IDs (*e.g.*, black upper clothes are similar to each other in appearance), which might be unhelpful for the Re-ID task. Next, we show some examples of the noisy attributes. In Figure 1(b), we show two examples of annotation errors. Here the bounding boxes for the persons may be distracted by the persons who are walking together, making the attribute annotation for the images transferred to the attribute values from the persons who are also in the image. In Figure 1(c), we show two examples of attribute inconsistency for ID-level attributes caused by object occlusions. In the left image, the person is occluded

Manuscript received December 24, 2019; revised May 21, 2020 and September 27, 2020; accepted October 18, 2020. Date of publication November 16, 2020; date of current version November 25, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0100704, in part by the Science and Technology Commission of Shanghai, China, under Grant 20511100300, in part by the National Natural Science Foundation of China under Grant 61902247, and in part by the Shanghai Sailing Program under Grant 19YF1424400. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Heng Tao Shen. *(Corresponding authors: Li Niu; Liqing Zhang.)*

The authors are with the MOE Key Laboratory of Artificial Intelligence, Department of Computer Science and Engineering, Shangha Jiao Tong University, Shanghai 200240, China (e-mail: c.sis@sjtu.edu.cn; ustcnewly@sjtu.edu.cn; zhang-lq@cs.sjtu.edu.cn).

Digital Object Identifier 10.1109/TIP.2020.3036762

1057-7149 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

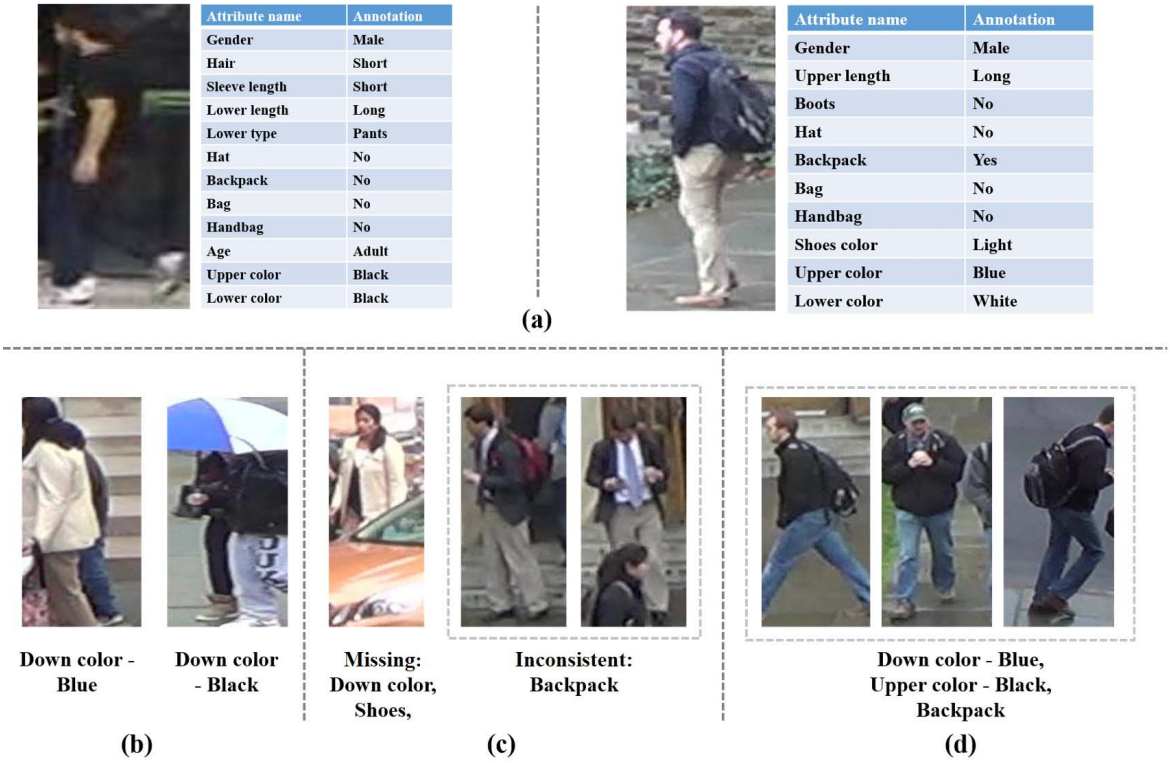


Fig. 1. (a) Example pedestrian images with attribute annotations, in which the images are from Market1501 [19] and DukeMTMC-reID [20] dataset with the annotations collected from [17]; (b) Examples of annotation errors; (c) Examples of attribute inconsistency for ID-level attributes; (d) Examples of non-discriminative attributes.

by a car making the attributes of lower body unavailable. In the right two images, the person turns around and makes the backpack attribute hard to recognize. In Figure 1(d), we show three images that all have black upper clothes and blue lower clothes. It is difficult to distinguish these three persons based on the attributes: upper clothes color and lower clothes color. Another reason for the attributes being non-discriminative is that the attributes are too rare or common. For example, on Market1501 dataset, fewer than 5% people are wearing hats, making the hats attribute biased and less helpful for Re-ID task. We refer to the abovementioned incorrectly labeled attributes or non-discriminative attributes as noisy attributes, while the other attributes are called key attributes.

To statistically analyze the noisy attributes, we relabel the attribute annotations to check the ratio of errors. Because there are no ground-truth image-level attribute annotations and manually annotating attributes is labor-intensive, we randomly select 1000 images from DukeMTMC-reID [20] dataset and relabel the attribute annotations for these images to make statistical analysis. The ratios of noisy annotations are shown in Table I. Since the definition of non-discriminative is very subjective, we only analyze the statistical results of 1) annotation errors and 2) attribute inconsistency. Please note that if an attribute is recognized as annotation error, then it will be excluded from inconsistent attributes. As we can see in Table I, some attributes (*e.g.*, boots, shoe color, upper color, lower color) have more than 10% of attribute errors, which may heavily affect the performance of Re-ID task. The average error ratio over all attributes is about 10% after combining 1) and 2).

TABLE I

STATISTICAL ANALYSIS OF ATTRIBUTE NOISE ISSUE ON DUKEMTMC-REID [20] DATASET. NUMBERS ARE REPORTED IN PERCENTAGE OF ERRORS

| Attributes | Annotation errors ratio (%) | Attribute inconsistency ratio (%) |
|---|---|---|
| Gender | 9.6 | 2.8 |
| Upper length | 0.5 | 0 |
| Boots | 3.7 | 11.9 |
| Hat | 0.9 | 3.2 |
| Backpack | 2.5 | 2.3 |
| Bag | 2.8 | 1.8 |
| Handbag | 2.9 | 2.3 |
| Shoes color | 6.4 | 12.4 |
| Upper color | 13.3 | 0 |
| Lower color | 11.5 | 5.0 |
| Average | 5.4 | 4.2 |

To address the attribute noise issue, we propose a novel Attribute Selection Module (ASM) embedded in AAB, in which reinforcement learning is used to select the key attributes by dropping the noisy attributes based on our designed reward mechanism. The reward mechanism will encourage ASM to drop the noisy attributes for higher retrieval accuracy. With the designed reward mechanism, our ASM can drop noisy attributes sequentially (one at each time step) and adaptively (determine the next one based on the current state). The remaining key attributes are expected to be more robust and discriminative. After dropping the noisy attributes, AAB only aggregates the attention related to key attributes. Our model achieves a better score for Re-ID evaluation by dropping the noisy attributes. Experiment results show the competitive performance on two datasets with attribute annotations, and

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

also one dataset without attribute annotations, which prove that our model can also work with the unsupervised attribute information.

The main contributions of our work are listed as follows:

- We propose a novel Attribute Attentional Block (AAB), which exploits fine-grained attribute attention to improve the performance of Re-ID task.
- We further propose a novel Attribute Selection Module (ASM) embedded in AAB, which uses reinforcement learning to drop the noisy attributes.
- We show that our method achieves the state-of-the-art performance on three popular Re-ID datasets. Compared with the baselines using attribute information, our method achieves as high as 3% improvement based on rank-1 accuracy, which demonstrates the advantage of our method in utilizing attribute information.

## II. Related Works

In this section, we provide a brief review of existing Re-ID methods from the following aspects.

### A. Re-ID Based on Deep Learning

There are mainly two research lines for the Re-ID task: the first one is representation learning [21]–[23], which learns a discriminative ID-level representation and then calculate feature distance between query and gallery images. While metric learning is to optimizes the distance metric [9], [24]–[27]. For example, to pull the images with the same ID together by minimizing the feature distance and push away the images with different IDs by maximizing the feature distance [24]. There are also methods maintaining a distance threshold for a selected anchor image, to keep difference between the distance from the anchor image to the images with the same ID and the distance from the anchor image to the images with different IDs larger than the threshold [26]. Currently, the research trend is combining representation learning and metric learning [13], [28]–[36] in one system. In this work, we also unify representation learning and metric learning in an end-to-end system.

### B. Re-ID Based on Attention Models

Re-ID models suffer from misalignment issue due to the pose variation, wrong bounding boxes, and variation of the camera views. To alleviate these problems, some methods try to match the invariant parts of the persons [37], [38]. PurifyNet [39] refines the wrong ID labels for noisy images generated by wrong bounding boxes and reuses these images to train the model with refined labels to boost Re-ID performance. The attention mechanism [40] can guide the network towards the discriminative local information. Some previous attention based approaches for the re-id task employ hard attention [31], [32], [41], [42] or soft attention [13], [14], [43] to alleviate the misalignment issue in the re-id task, which is caused by different person poses and angles of the camera view. Nevertheless, none of these methods leverage attribute information, which is proved to be favorable for Re-ID [17], [44]. More recently, Tay *et al.* [18] proposed a model based on attribute information and Class Activation Map (CAM) [45], which can generate attribute attention map. However, CAM is post-hoc attention mechanism based on classification activation map and ill-suited to alleviate misalignment issue, while we use a trainable attention mechanism as in [14].

### C. Re-ID Based on Attribute

The attribute can be categorized into ID-level attribute and image-level attribute, as discussed in Section I. For image-level attribute, some previous works use the attribute information to select tuples for metric learning [22], [44] or mine the rules for verifying persons [46]. There are more works using ID-level attribute. For example, Lin *et al.* [17] proposed to combine person identification and ID-level attribute recognition while Schaul *et al.* [47] proposed to use the attribute information as auxiliary information to train a network with triplet loss. Tay *et al.* [18] combine body part and attribute information with attention modules. These methods do not identify the key attributes and thus suffer from attribute annotation noise. In contrast, our auxiliary task can select key attributes and only incorporate useful attribute information into the Re-ID task.

### D. Re-ID Based on Reinforcement Learning

Reinforcement Learning (RL) trains an agent to exploit the experience during exploration and simultaneously maximize the cumulative reward. There are only a few works using RL for person Re-ID. For instance, Lan *et al.* [41] proposed a method to refine the bounding box of persons to improve Re-ID task performance by RL. In another work [48], RL is applied to learn the adequate number of image pairs to verify in the multi-shot Re-ID task. Distinguished from the above two methods, our proposed ASM module uses reinforcement learning to select key attributes in the auxiliary task, which has not been studied before.

## III. Our Proposed Method

### A. Overview

Given a dataset composed of pedestrian images (each image contains a detected bounding box enclosing a pedestrian) from different persons captured by various surveillance cameras, we split the dataset into a training set and test set according to person ID, so that training IDs and test IDs have no overlap. Each training ID is associated with a suite of ID-level attribute labels. In the testing stage, the test set is divided into a set of query images and a set of gallery images. Then, given a query image, the goal of person Re-ID is to rank all the gallery images according to the probability that it belongs to the same person as the query image. Formally, we assume the dataset contains an image set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ with ID labels $\mathbf{Y} = \{y_1, \ldots, y_K\}$ where $y_i \in \{1, \ldots, N\}$, in which $K$ is the number of images and $N$ is the number of persons. Each training image $\mathbf{x_i}$ from training IDs has the ID-level attribute labels $\mathbf{z}_i = (z_i^1, \ldots, z_i^M) \in \mathcal{N}^M$, in which $M$ is the number of attributes.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

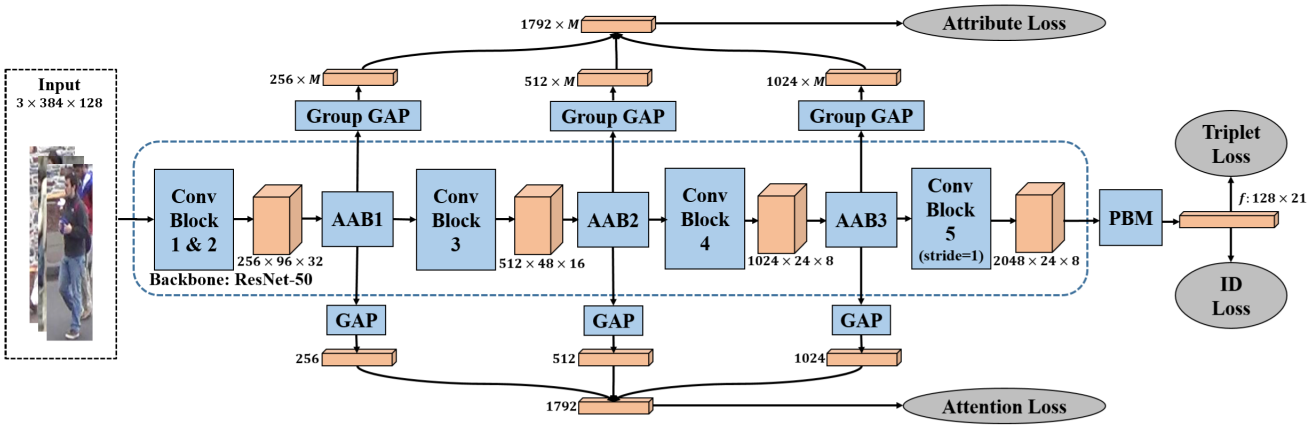


Fig. 2. The pipeline of our proposed method. The backbone network is ResNet50. There are three AABs added after the second, third and fourth Conv Block of ResNet50, supervised by attribute loss and attention loss. We also apply ID loss and triplet loss to the final image embedding. The detailed structure of AAB is shown in Figure 3 and described in Section III-B.

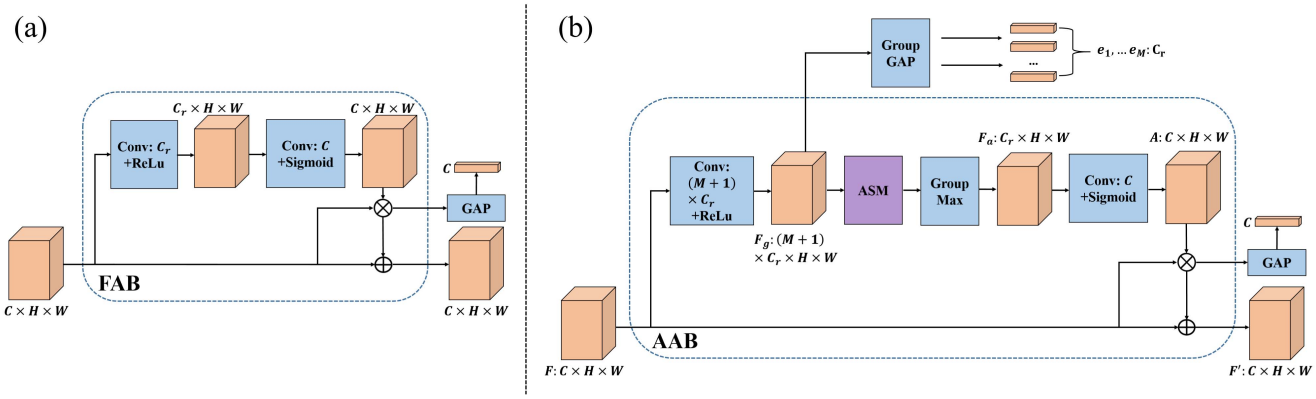


Fig. 3. The comparison between (a) Fully Attentional Block (FAB) [14] and (b) our proposed Attribute Attentional Block (AAB).

As depicted in Figure 2, given a batch of images, we feed them through the same CNN model, *i.e.*, ResNet50 [12] pretrained on ImageNet [10], which has five convolution blocks. We use the feature maps of the second, the third and the fourth convolution block to generate attention masks by our proposed Attribute Attentional Block (AAB), which are multiplied with the feature maps and added back to the mainstream of the network. AAB learning is guided by ID information with attention loss and attribute information with attribute loss, to focus on the entire human body and the local information related to $M$ attributes. The details of AAB can be found in Section III-B. Furthermore, to select key attributes, we design an Attribute Selection Module (ASM) and put it in the third AAB module. For the details of ASM, please refer to Section III-C. Then, the last convolution block is used to generate the final feature map. After that, Pyramidal Branches Model (PBM) [34] is applied to generate part-based features of the image, which are pooled to the final image embedding $\mathbf{f}$ and fed into the most commonly used Re-ID losses: triplet loss and ID loss. In Section III-D we will demonstrate how to combine all the modules and objective functions to train and evaluate our proposed model.

### *B. Attribute Attentional Block (AAB)*

Misalignment problem, which is caused by various person poses and different camera viewpoints, is massively raised in Re-ID task. Attention model [13], [14], [49] is a popular way to deal with this issue by learning regions of interest and extract features from those regions. One example is Fully Attentional Block (FAB) proposed in [14]. As illustrated in Figure 3, FAB uses a $1 \times 1$ convolution layer with ReLU activation followed by another $1 \times 1$ convolution layer with sigmoid activation to generate attention map. Note that FAB creates both spatial and channel-wise attention together, producing an attention tensor. In this way, FAB can localize discriminative local features and tackle the misalignment problem [14].

Inspired by FAB [14], we propose an Attribute Attentional Block (AAB), which utilizes attribute information to obtain fine-grained attribute attention. Following [14], we apply our proposed AAB module after the second, the third and fourth block of the ResNet50 backbone. The comparison between FAB and our AAB is shown in Figure 3. With an input feature tensor $\mathbf{F}$ with size $C \times H \times W$, we apply a $1 \times 1$ convolution layer with $(M+1) \times C_r$ outputs and a ReLU layer to generate the grouped attribute feature $\mathbf{F}_g$ with size $(M+1) \times C_r \times H \times W$, in which we set $C_r = C/16$ to reduce the number of parameters. We split $\mathbf{F}_g$ evenly into $\mathbf{F}_g^0, \ldots, \mathbf{F}_g^M$, where $\mathbf{F}_g^0$ represents global bottleneck feature and $\mathbf{F}_g^1, \ldots, \mathbf{F}_g^M$ stand for $M$ attribute bottleneck feature. Then $\mathbf{F}_g$ is fed into two branches: attribute branch and global branch.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

*1) Attribute Branch:* We apply Global Average Pooling (GAP) to each $\mathbf{F}_g^i$, named group GAP, and generate $M+1$ different embeddings $\mathbf{e}^0, \ldots, \mathbf{e}^M$ with length $C_r$. Then $\mathbf{e}^1, \ldots, \mathbf{e}^M$ are supervised by attribute classification loss. Specifically, we concatenate all the attribute embeddings from all the three AAB modules (see Figure 2) and perform attribute classification based on the concatenated attribute embeddings. Recall that $z^i$ is the attribute label of the $i$-th attribute as mentioned in Section III-A. By using $FC_j(\cdot)$ to denote $j$-th output of a Fully-Connected layer, the attribute loss can be defined as:

$$L_{attr} = \sum_{i=1}^{M} -log\frac{exp(FC_{z^i}(\mathbf{e}^i))}{\sum_j exp(FC_j(\mathbf{e}^i))}. \tag{1}$$

*2) Global Branch:* Since noisy attributes may harm the Re-ID performance as discussed in Section I, we design an Attribute Section Module (ASM) to drop noisy attributes. Our ASM attempts to select the attribute bottleneck features corresponding to the key attributes from $\mathbf{F}_g^1, \ldots, \mathbf{F}_g^M$, and put them together with the global feature bottleneck $\mathbf{F}_g^0$. In particular, ASM will generate a binary mask

$$\mathbf{b} = (1, b^1, \ldots, b^M), \tag{2}$$

where $b^i \in \{0, 1\}$. $b^i = 1$ (*resp.*, $b^i = 0$) means we keep (*resp.*, remove) $\mathbf{F}_g^i$ to generate attention. We then apply element-wise max to the masked attribute bottleneck features $\{b^i \mathbf{F}_g^i |_{i=0}^M\}$ and get an aggregated attribute feature $\mathbf{F}_a$:

$$\mathbf{F}_a = max(\mathbf{F}_g^0, b^1\mathbf{F}_g^1, \ldots, b^M\mathbf{F}_g^M). \tag{3}$$

$\mathbf{F}_a$ will be a feature tensor with size $C_r \times H \times W$. After that, a $1 \times 1$ convolution layer will transform $\mathbf{F}_a$ back to the same size as input feature tensor $\mathbf{F}$. Finally, sigmoid function is applied to make the attention scores range in $[0, 1]$, leading to the aggregated attribute attention $\mathbf{A}$.

The aggregated attribute attention $\mathbf{A}$ is supervised by ID classification loss, which will guide AAB to learn better attention information:

$$L_{atte} = -log\frac{exp\left(FC_y(GAP(\mathbf{F} * \mathbf{A}))\right)}{\sum_j exp\left(FC_j(GAP(\mathbf{F} * \mathbf{A}))\right)}, \tag{4}$$

where $*$ stands for element-wise multiplication, $y$ is the ID label of input image, and $GAP$ means global average pooling.

Overall, after we apply AAB to the main branch of the network, we will change the input feature tensor $\mathbf{F}$ to $\mathbf{F}'$:

$$\mathbf{F}' = \mathbf{F} * \mathbf{A} + \mathbf{F}. \tag{5}$$

Here, we choose to apply residual attention instead of direct attention (*i.e.*, $\mathbf{F}' = \mathbf{F} * \mathbf{A}$) or residual feature (*i.e.*, $\mathbf{F} = \mathbf{F} + \mathbf{A}$) due to the following reasons. Previous works [13], [14] have shown that attention can alleviate the misalignment problem in Re-ID task. Moreover, the potential of residual attention [50] in improving the Re-ID performance has been demonstrated in [14]. Hence, we opt for residual attention in our method, but we also investigate different choices in Section IV-C.

Finally, we pass $\mathbf{F}'$ through Pyramidal Branches Model (PBM) [34] and generate a 128-dim part-based image embedding for each of 21 parts, which are concatenated as a long vector $\mathbf{f}$. We use $\mathbf{f}$ as the final image embedding.

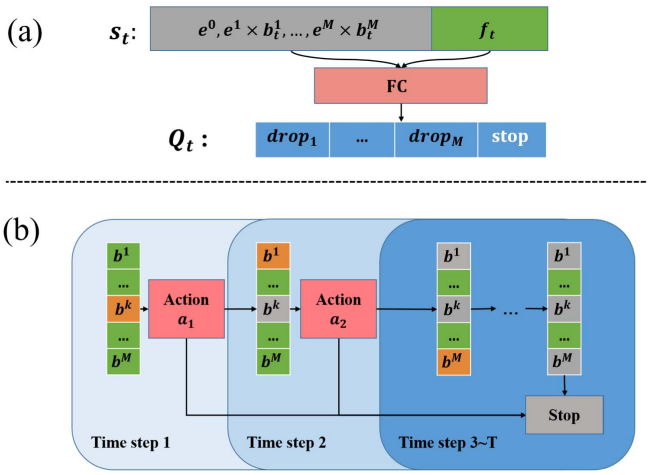


Fig. 4. An illustration of the proposed Attention Selection Module (ASM): (a) The structure of Q-network with state $s_t$ and Q-values for each type of action. (b) The process of updating $b_t$ after taking a sequence of actions in an episode, in which the orange entry indicates the selected attribute at each time step $t$ and the grey entry indicate the attribute is already dropped (*i.e.*, $b^k = 0$).

### C. Attribute Selection Module (ASM)

Recall that in Section III-B, an Attribute Selection Module (ASM) is embedded in AAB to select key attributes. Prior to the introduction of ASM, there are two trivial alternative strategies to get binary mask $\mathbf{b}$ in AAB:

- *All*: Always keep all the attributes, *i.e.*, $b^i = 1, \forall i$.
- *Random*: Select or drop the attributes half-and-half, *i.e.*, each $b^i$ is randomly set to 0 with probability 0.5.

Note that for the first two AAB block, we apply *All* strategy because we are not willing to drop information in the early stages in the network. We only use ASM to select key attributes in the third AAB block (AAB3 in Figure 2).

The goal of ASM is to determine the key attributes of each image to facilitate the Re-ID task. A naive way is comparing all the possible selections with $2^M$ different choices, but it is too slow because we have to enumerate all the different $2^M$ choices and calculate the outcome to make decisions. Alternatively, we can add attributes to an empty set one by one or drop attributes from the full set one by one. Considering useful attributes should be more than noisy attributes, we choose to drop noisy attributes one by one from the full set.

In the remainder of this paper, we refer to the whole process of dropping noisy attributes as an episode. At each time step in an episode, we recognize one noisy attribute and drop it by $b^i = 0$. We formulate this decision making process under the reinforcement learning framework, in which an agent needs to interact with the dynamic environment by exploring the unknown and exploiting the experience. The policy for an agent to take the optimal *actions* on the premise of different *states* is learnt by maximizing the cumulative *rewards* in each episode. We train the agent to learn the optimal policy via Q-learning [51] by optimizing a Q-network. The Q-network takes *state* as input and outputs Q-value, which indicates the *action* to take. In our problem, *state*, *action*, and *reward* are defined as follows.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

*1) State:* The state describes the observation of the agent in the environment. By using $s_t$ to represent the state at the time step $t$, the state

$$\mathbf{s}_t = (\mathbf{e}^0, \mathbf{e}^1 \times b_t^1, \ldots, \mathbf{e}^M \times b_t^M, \mathbf{f}_t) \tag{6}$$

consists of $\mathbf{e}^i$, $b_t^i$ and $\mathbf{f}_t$ (see Figure 4(a)), in which $b_t^i$ (*resp.*, $\mathbf{f}_t$) is the binary mask (*resp.*, the final image embedding) at time $t$ and $\mathbf{e}^i$ is the attribute embedding. We concatenate all these embeddings and use Q-network (*i.e.*, two FC layers) to generate $M+1$ outputs with each output corresponding to the Q-value for each action.

*2) Action:* At each time step $t$ in an episode, the agent needs to know whether to drop one noisy attribute or terminate the current episode. Therefore, the agent has in total $M+1$ types of actions, including the "stop" action and "drop" action for each attribute. Formally, we use $a_t$ to denote the action at the time step $t$. Given an image, at the time step $t$, the agent selects the $i$-th attribute as the noisy attribute to drop (*i.e.*, $a_t = i$), or chooses "stop" (*i.e.*, $a_t = M+1$) to terminate the current episode. When $t > 1$ and $a_t \neq M+1$, the binary mask $b_t$ will be updated as follows: $b_{t+1}^i = 0$ if $a_t = i$ (see Figure 4(b)). Note that the agent is not allowed to choose the attribute that has been dropped before in the same episode (*i.e.*, $b_t^i = 0$), to avoid the duplicated dropping of key attributes. In an extreme case, at the time step $M+1$, the agent has dropped all $M$ attributes and will be enforced to choose "stop".

*3) Reward:* The reward is defined as the benefit or penalty that the agents will receive after taking specific action in the current state. If $a_t = M+1$, no rewards are given to the agent, so we only discuss the rewards for the time steps that $a_t \neq M+1$. For each input image $\mathbf{x}$ with ID label $y$, we sample $n \times k$ *reference images* consisting of $n$ different images from each of $k$ different persons. We set $n = 4$ and $k = 16$ in experiments and make sure that ID label $y$ is among the $k$ person IDs. Hence, there will be $n$ reference images with the same ID label as $\mathbf{x}$, which are referred to as positive images.

By simulating the query-gallery retrieval process with $\mathbf{x}$ (*i.e.*, the query image) and $n \times k$ reference images (*i.e.*, the gallery images), we expect the agent to improve mAP and ranking accuracy with the help of our designed rewards. We compute the Euclidean distance between reference images and query image $\mathbf{x}$ based on their image embeddings $\mathbf{f}$, and rank reference images in increasing order according to their distances to $\mathbf{x}$. Note that we perform *Random* strategy for the reference images to calculate their image embeddings instead of *All* based on the experimental comparison. We sort $n$ rank positions of positive images into a rank tuple:

$$\mathbf{p}_t = (p_t^1, \ldots, p_t^n), \tag{7}$$

where $p_t^1, \ldots, p_t^n$ are in an increasing order. Then, the reward $r_t$ is calculated by the summation of the following two terms:

- **AP reward:** Average Precision (AP) at time $t$ is calculated as

$$AP_t = \frac{1}{n} \sum_{i=1}^{n} \frac{i}{p_t^i}. \tag{8}$$

AP is a common metric for the Re-ID task. We hope that the AP score could be improved after the agent drops a noisy attribute. Hence, we define AP reward as

$$r_t^{AP} = AP_{t+1} - AP_t. \tag{9}$$

- **Ranking reward:** Ranking position is another important metric for the Re-ID task. We expect ranking positions of positive images could be advanced after the agent drops a noisy attribute, so the ranking reward is defined as

$$r_t^{RR} = \begin{cases} +1, & if\ \exists i, p_{t+1}^i < p_t^i \wedge \forall j < i, \\ & p_{t+1}^j = p_t^j, \\ -1. & otherwise. \end{cases} \tag{10}$$

*4) Training the Agent::* Based on the definitions of state, action, and rewards, we train the agent to take actions in the given state to maximize the cumulative reward. In particular, we adopt standard deep Q-learning [51], which estimates off-policy Q-value to maximize the expected value of the total reward for each episode. we use deep Q-learning [51] to train the agent to seek for the optimal policy. We use $Q(s_t, a_t)$ to represent the $Q$-value at the time step $t$, which estimates the discounted cumulative reward from the current time step to the end of the episode. $Q(s_t, a_t)$ can be calculated based on the recursive formulation

$$Q(s_t, a_t) = r_t + \gamma \max_{a'} Q(s_{t+1}, a'), \tag{11}$$

where $\gamma \in (0, 1)$ is called discount factor, that is, larger $\gamma$ makes the agent pay more attention to the future accumulated reward while smaller $\gamma$ makes the agent pay more attention to the instant reward (*i.e.*, $r_t$). We use a simple network with one FC layer to estimate the Q-value $Q(s_t, a_t)$ given $s_t$ and $a_t$, which is referred to as Q-network. Instead of feeding $s_t$ and each type of action $a_t$ into Q-learning network, we use a single forward pass to compute Q-values for all actions, as shown in the left subfigure in Fig. 3(a). Since there are a total of $M+1$ types of actions, the output size of our Q-network is $M+1$.

By using $\theta$ to denote the model parameters of Q-network and the modules preceding Q-network, we adopt $\epsilon$-greedy learning [52] to update $\theta$ in the training stage. In each training iteration, we randomly sample a batch of image pairs and create a sequence of transitions for each image pair. Specifically, in the $i$-th training iteration, with the model parameters $\theta_{i-1}$ and the initial state $s_1$, we select actions according to $\epsilon$-greedy policy [52] and obtain the corresponding reward until one episode terminates, yielding a sequence of transitions $\{(s_t, a_t, r_t, s_{t+1})\}$. The $\epsilon$-greedy policy means that at each time step $t$; the agent selects the action corresponding to the maximum Q-value with probability $1-\epsilon$ and a random action with probability $\epsilon$, from the available set of actions whose associated attributes have not been selected yet in the same episode. For each transition $(s_t, a_t, r_t, s_{t+1})$, we can estimate the optimal Q-value $Q^*(s_t, a_t)$:

$$Q^*(s_t, a_t) = \begin{cases} r_t & if\ a_t = M+1, \\ r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta_{i-1}) & otherwise. \end{cases} \tag{12}$$

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

**Algorithm 1** Training Based on the ASM Loss

**Input:** Training images with their ID labels and the attributes of training IDs.
**Output:** Model parameters $\theta$

```
Initialize θ_0.
for i ← 1, 2, ..., I do
    Sample a batch of images B and initialize the transition memory D = ∅.
    for x_i in B do
        Generate transitions {(s_t, a_t, r_t, s_{t+1})} within one episode.
        D = D ∪ {(s_t, a_t, r_t, s_{t+1})}.
    end for
    for (s_t, a_t, r_t, s_{t+1}) in D do
        if a_t = M + 1 then Q*(s_t, a_t) ← r_t.
        else Q*(s_t, a_t) ← r_t + γ max_{a'} Q(s_{t+1}, a'; θ_{i-1}).
        end if
    end for
    Update θ_i with gradient ∇_{θ_i} E_{s_t,a_t}[(Q(s_t, a_t; θ_i) − Q*(s_t, a_t))^2].
end for
return θ_I
```

We update model parameters $\theta_i$ to enforce $Q(s_t, a_t; \theta_i)$ to be close to the optimal Q-value $Q^*(s_t, a_t)$.The training algorithm with the ASM loss is summarized in Algorithm 1, in which $\gamma$ is set as 0.9, $\epsilon$ shrinks from 1 to 0.1 linearly in the first 200000 iterations. We use a replay buffer with size 10000 and Double-DQN [53] to help the agent learn better policy.

We update ASM model parameters $\theta_i$ to enforce $Q(s_t, a_t; \theta_i)$ to be close to the optimal Q-value $Q^*(s_t, a_t)$, leading to our ASM loss:

$$L_{asm} = \mathbb{E}_{s_t,a_t}[(Q(s_t, a_t; \theta_i) - Q^*(s_t, a_t))^2]. \quad (13)$$

### D. System Configuration

Except for the three losses defined in Eqn. (13), (1) and (4), we employ another two commonly used Re-ID losses to guide the model training:

*1) ID Loss:* ID loss is the same as cross-entropy ID classification loss defined as:

$$L_{id} = \sum_{i=1}^{K} -log\frac{exp(FC_{y_i}(\mathbf{f}_i))}{\sum_j exp(FC_j(\mathbf{f}_i))}, \quad (14)$$

where $\mathbf{f}$ is the final image embedding and $FC(\cdot)$ is an fully-connected layer applied after $\mathbf{f}$.

*2) Triplet Loss:* Given a triplet of samples $(\mathbf{x}, \mathbf{x}_+, \mathbf{x}_-)$ with ID labels $(y, y_+, y_-)$, which satisfy $y = y_+$ and $y \neq y_-$, triplet loss will push apart $\mathbf{f}$ and $\mathbf{f}_-$ while pulling together $\mathbf{f}$ and $\mathbf{f}_+$. Triplet loss can be defined as:

$$L_{tri} = \sum_{\mathbf{x},\mathbf{x}_+,\mathbf{x}_-} max(d(\mathbf{f}, \mathbf{f}_+) - d(\mathbf{f}, \mathbf{f}_-) + \delta, 0), \quad (15)$$

where $d(\cdot, \cdot)$ means Euclidean distance and $\delta$ is the margin that is set as 1.2 in our experiments.

Overall, the full objective function of our proposed model is:

$$L = L_{ID} + L_{tri} + \lambda_{attr}L_{attr} + \lambda_{atte}L_{atte} + \lambda_{asm}L_{asm}, \quad (16)$$

where $\lambda_{attr}$, $\lambda_{atte}$, and $\lambda_{asm}$ are trade-off parameters. We empirically set $\lambda_{attr} = 0.25$, $\lambda_{atte} = 0.2$, and $\lambda_{asm} = 1$.

*3) Stabilizing ASM Training:* In practice, we observe that joint training of ASM and attention modules is unstable and time-consuming. Therefore, we pretrain AAB without ASM to stabilize subsequent ASM training. We train our proposed model in the following three stages, and the details can be found in Section IV-B.

- **Stage1:** Train the whole model using *Random* strategy for AAB3. ASM is not optimized, which means $\lambda_{ASM} = 0$;
- **Stage2:** Train ASM while freezing all the network parameters except for ASM;
- **Stage3:** Train the whole model freezing first four convolution blocks and all the AAB blocks.

In the testing stage, given a query image and all gallery images, we obtain their final image embedding $\mathbf{f}$ and calculate Euclidean distance. ASM is applied in the testing stage to drop the noisy attribute features. Then, we rank all gallery images based on their distance to the query image.

### E. Difference Between Other Related Methods

FAB [14] only pays attention to person ID information and does not consider the attribute information. In contrast, our AAB learns M attribute bottleneck features and one global bottleneck feature under the supervision of attribute information. Then, bottleneck features are aggregated via group max to generate attribute-enhanced full attention map, which is better than the full attention map obtained by FAB without utilizing attribute information. Besides, a novel ASM is embedded in AAB to further boost the performance by dropping noisy attributes.

CA3Net [54] and AANet [18] only extract the attribute feature based on attribute attention map for final attribute recognition or person Re-ID, in which the attribute feature is obtained in the last stage of the network. In contrast, our AAB can learn a full attention map in different stages of the network, which can alleviate the misalignment issue (see Section III-B) in a similar way to FAB. Also, CA3Net and AANet do not consider the attribute noise issue. In contrast, our AAB utilizes ASM to drop noisy attributes sequentially and adaptively.

## IV. Experiments

In this section, we compare our method with our special cases and state-of-the-art baselines on three benchmark datasets together with in-depth qualitative analyses.

### A. Datasets

We evaluate our method and baselines on three datasets: Market-1501 [19], DukeMCMT-ReID [20], and CUHK03 [24]. The detailed setting for each dataset can be found in Supplementary. Here, we will demonstrate how we get the attribute annotations for each dataset.

*1) Market-1501:* Market-1501 [19] dataset contains 32668 images of 1501 IDs with annotated bounding boxes are detected and collected using the Deformable Part Model (DPM) pedestrian detector. View overlapping

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

exists among different cameras, including 5 high-resolution cameras, and one low-resolution camera. The dataset into a training set with 12936 images of 751 persons and a testing set of 750 persons containing 3368 query images and 19732 gallery images. For the Market-1501 [19] dataset, we train our model with annotated ID-level attribute labels for 30 attributes provided in [17]. We merge 9 attributes for lower-body clothing color, 8 attributes for upper-body clothing color, and 4 attributes for age, leading to in total 12 attributes.

*2) DukeMCMT-ReID:* The DukeMTMC-reID [20] dataset is a subset of the DukeMTMC dataset specifically collected for person re-id. 1404 identities appear in more than two cameras while 408 (distractor) identities appear in only one camera. The training set consists of 16522 images with 702 identities and a testing set which consists of 2228 query images of the other 702 identities and 17661 gallery images of 702 identities plus 408 distractor identities. We train our model with annotated ID-level attribute labels for 23 attributes provided in [17] for the DukeMTMC-reID [20] dataset. Similar to the Market-1501 dataset, we merge 8 attributes for lower-body clothing color and 7 attributes for upper-body clothing color, resulting in totally 10 attributes.

*3) CUHK03:* Following [55], we split all IDs into 767 training IDs and 700 test IDs. From each camera, one image is selected as the query for each ID and the rest of images are used to construct the gallery set. There are two ways of annotating bounding box, including labeled by human or detected by a detector. In this paper, we are using the detected bounding boxes. The dataset consists of 7365 training, 1400 query, and 5332 gallery images. Note that CUHK03 dataset does not have attribute annotations, so we explore discovering latent attributes in an unsupervised manner. Without loss of generality, we assume that there are 10 latent attributes ($M = 10$). The training process is the same as the other two datasets except that $\lambda_{attr} = 0$ in Eqn. (16) due to the absence of attribute supervision. In this case, our ASM can also learn to drop noisy latent attributes.

### B. Implementation Details

Recall that we have three training stages. We train each stage for 400000 iterations. The initial learning rate is 0.001 for the first two stages and 0.0001 for the last stage. In each stage, after 200000 and 300000 iterations, the learning rate will be discounted by 0.1. The training batch size is 64, including 4 different images from 16 different IDs. Random flip and random erasing augmentation [56] are applied to the input images. Hard example mining is used for triplet loss

We adopt two metrics for evaluation: top-1 accuracy and mean Average Precision (mAP) which are commonly used for Re-ID evaluation.

### C. Ablation Studies

To evaluate our proposed AAB and ASM modules, we conduct extensive ablation studies in this section. We evaluate 21 special cases including our full-fledged method on DukeMTMC-reID dataset and results are listed in Table II.

TABLE II
ABLATION STUDY ON DUKEMTMC-REID. THE BEST RESULTS IN EACH STAGE ARE IN BOLDFACE

| Stages | Settings | mAP(%) | R1(%) |
|---|---|---|---|
| Stage 1 | w/o AAB, w/o PBM | 66.6 | 80.2 |
| | w/o AAB, PBM→PCB | 73.6 | 84.4 |
| | w/o AAB | 76.3 | 86.5 |
| | AAB→FAB | 76.5 | 86.6 |
| | $\mathbf{F}+\mathbf{F}*\mathbf{A}\rightarrow\mathbf{F}+\mathbf{A}$ | 78.2 | 87.1 |
| | $\mathbf{F}+\mathbf{F}*\mathbf{A}\rightarrow\mathbf{F}*\mathbf{A}$ | 78.4 | 87.5 |
| | *random→All* | 77.8 | 87.3 |
| | Average | 78.3 | 87.9 |
| | Late-fusion | 78.6 | 87.9 |
| | Ours | **78.7** | 88.0 |
| | Double | 78.6 | **88.2** |
| Stage 1+2 | *random→All* (ref) | 78.6 | 88.9 |
| | Threshold | 78.5 | 88.0 |
| | Regression | 79.0 | 88.3 |
| | Ours | **79.8** | **89.2** |
| Stage 1+2+3 | w/o AAB1 | 79.8 | 89.2 |
| | w/o AAB1&2 | 78.9 | 88.5 |
| | w/o PBM | 77.1 | 87.5 |
| | PBM→PCB | 78.7 | 88.5 |
| | Ours ($\lambda_{attr} = 0$) | 79.0 | 88.5 |
| | Ours | **80.4** | **89.9** |

These special cases can be divided into three groups based on the stages used to train the models (see Section III-D). "Stage 1" means the models are only trained with Stage 1, "Stage $1+2$" means the models are trained with first two stages, and "Stage $1+2+3$" means the models are trained with all the three stages.

*1) Stage 1:* To evaluate our proposed AAB, we remove AAB ("w/o AAB") or replace AAB with FAB ("AAB→FAB"). The obtained results are worse than "Ours", which corroborates the effectiveness of our AAB.

We further remove Pyramidal Branches Model (PBM) [34] (see Figure 2) or replace PBM with Part-based Convolutional Baseline (PCB) proposed in [33]. The obtained results "w/o AAB, w/o PBM" and "w/o AAB, PBM→PCB" are worse than "w/o AAB", which shows the necessity of using PBM.

Next, we explore two variants of AAB. We use $\mathbf{F}' = \mathbf{F}+\mathbf{A}$ (*resp.*, $\mathbf{F}' = \mathbf{F}*\mathbf{A}$) to replace Eqn. (5), which is referred to as "$\mathbf{F}+\mathbf{F}*\mathbf{A}\rightarrow\mathbf{F}+\mathbf{A}$" (*resp.*, "$\mathbf{F}+\mathbf{F}*\mathbf{A}\rightarrow\mathbf{F}*\mathbf{A}$"). We observe that two variants are inferior to "Ours", implying the advantage of residual residual attention.

Note that we use *Random* strategy for AAB3 in Stage 1, another choice is *All* strategy (see Section III-D). It can be seen that *All* strategy is less effective than *Random* strategy by comparing "random→all" with "Ours".

We use group "max" as the strategy to aggregate attribute bottleneck features in Equation 3. Here we alternate the strategy by using "average" and "late-fusion" respectively, in which "max" shows the best performance.

To justify that the performance improvement is not caused by insufficient training of the first stage, we train another 2e5 iterations for our Stage1 model. Obtained results

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

TABLE III
RESULTS OF OUR PROPOSED METHOD AND STATE-OF-THE-ART BASELINES ON THREE DATASETS. THE BEST RESULTS ARE IN BOLDFACE

| Method | Market-1501 | | DukeMTMC-reID | | CUHK03 | |
|---|---|---|---|---|---|---|
| | mAP(%) | R1(%) | mAP(%) | R1(%) | mAP(%) | R1(%) |
| PL-Net [57] | 69.3 | 88.2 | - | - | - | - |
| DHA-Net [58] | 76.0 | 91.3 | 64.1 | 81.3 | - | - |
| Mancs [14] | 82.3 | 93.1 | 71.8 | 84.9 | 60.5 | 65.5 |
| PCB+RPP [33] | 81.6 | 93.8 | 69.2 | 83.3 | 57.5 | 63.7 |
| LITM [59] | 83.9 | 93.9 | 74.5 | 85.9 | 69.1 | 71.8 |
| HPM [60] | 82.7 | 94.2 | 74.3 | 86.6 | 57.5 | 63.9 |
| DG-Net [61] | 86.0 | 94.8 | 74.8 | 86.6 | - | - |
| CASN [49] | 82.8 | 94.4 | 73.7 | 87.7 | 68.0 | 73.7 |
| Pyramid [34] | 88.2 | 95.7 | 79.0 | 89.0 | 74.8 | 78.9 |
| ABD-Net [15] | 88.3 | 95.6 | 78.6 | 89.0 | - | - |
| APR [17] | 66.9 | 87.0 | 55.6 | 73.9 | - | - |
| AANet [18] | 82.5 | 93.9 | 72.6 | 86.4 | - | - |
| CA3Net [54] | 80.0 | 93.2 | 70.2 | 84.6 | - | - |
| Ours | **88.6** | **96.1** | **80.4** | **89.9** | **78.0** | **80.2** |

are (78.6, 88.2), which are close to the reported results (78.7, 88.0) in Table II, showing that the model without using ASM has been fully trained in Stage 1.

*2) Stage* $1+2$*:* Recall that we apply *Random* strategy to the reference images when calculating rewards in our ASM (see Section III-C). An alternative choice is applying *All* strategy to the reference images ("random→all") and the results indicate that *Random* strategy is a better choice.

Our proposed ASM uses reinforcement learning to drop the noisy attributes, which could sequentially and adaptively select the discriminative attributes and drop the noisy attributes for Re-ID task. To justify the power of combining AAB and ASM with reinforcement learning, we design another two strategies "Regression" and "Threshold" to drop noisy attributes, which do not use reinforcement learning.

For "Regression", we replace group max by adding a regression layer to calculate weights for different attributes and multiplying these weights to the corresponding attribute bottleneck features. Then, the model is trained using weighted aggregated features with large weights (*resp.*, smaller weights) indicating clean attributes (*resp.*, noisy attributes)..

For "Threshold", we drop the attributes with low attribute prediction confidence. As we claimed in Section I, we categorize the noisy attributes into three different types: annotation errors, attribute inconsistency, and non-discriminative. The first two types could be directly derived from the attribute prediction confidence (similar to Eqn. 1, the confidence for the $i$-th attribute with attribute label $z^i$ for embedding $e^i$ is defined as $\frac{exp(FC_{z^i}(\mathbf{e}^i))}{\sum_k exp(FC_k(\mathbf{e}^i))}$) while the last one (non-discriminative) should be related to Re-ID task. As a reference experiment, we drop the attributes whose attribute prediction confidence is lower than the threshold 0.9, which is determined by cross-validation.

We compare the performance of "Regression", "Threshold" with our method in Table II. Comparing all three methods, we can see that "Threshold" method does not bring improvement for Re-ID task (compared with Stage1 "Ours"), which may be due to the lack of considering the effects of non-discriminative attributes. Compared with our method, "Regression" method cannot drop noisy attributes sequentially and adaptively. Moreover, a simple regression model is not specifically designed for denoising attributes, and the lower weights may not be reliable in indicating noisy attributes.

*3) Stage* $1+2+3$*:* To investigate the number of required AAB blocks, we report the results of "w/o AAB1" and "w/o AAB1&2", which are worse than "Ours". Thus, it is helpful to use all three AAB blocks. "Ours ($\lambda_{attr}=0$)" means training the model with $\lambda_{attr}=0$ in Eqn. (16) through all three stages. It can be seen that the performance is worse compared with "Ours", which verifies the benefit of attribute information. However, "Ours ($\lambda_{attr}=0$)" is still better than "w/o AAB" in "Stage 1", which means that our AAB can discover latent attributes and help improve performance even without attribute supervision.

Finally, we report the results by removing PBM or replacing PBM with PCB in our method. By comparing "w/o PBM" (*resp.*, "PBM→PCB") in "Stage $1+2+3$" with "w/o AAB, w/o PBM" (*resp.*, "w/o AAB, PBM→PCB") in "Stage 1", we can see that our proposed AAB can improve the performance on the foundation of different baseline frameworks.

### *D. Comparison With State-of-the-Art Methods*

We compare our method with the state-of-the-art baselines on three benchmark datasets. The baselines can be divided into two groups:

(1) AANet [18], APR [17] and CA3Net [54] using ID-level attribute information, which is close to our method;
(2) PL-Net [57], DHA-Net [58], DG-Net [61], CASN [49], Mancs [14], PCB + RPP [33], LITM [59], HPM [60], Pyramid [34], ABD-Net [15] which are currently state-of-the-art methods without using attribute information.

For fair comparison, all the results are based on single query setting with ResNet50 as the backbone network. Re-ranking process which can improve mAP is not used.

The results of different methods are reported in Table III. It can be seen that compared with the baselines using

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

NUMBER OF PARAMETERS (IN MEGA-BYTES) AND PERFORMANCE FOR DIFFERENT MODELS. RESULTS OBTAINED ON DUKEMTMC-REID

| Method | mAP(%) | R1(%) | #Parameters(M) |
|---|---|---|---|
| w/o AAB, w/o PBM | 66.6 | 80.2 | 23.5 |
| w/o AAB | 76.3 | 86.5 | 26.7 |
| w/o ASM | 78.7 | 88.0 | 30.2 |
| APR [17] | 55.6 | 73.9 | 24.4 |
| AANet [18] | 72.6 | 86.4 | 29.9 |
| CA3Net [54] | 70.2 | 84.6 | 31.0 |
| Ours | 80.4 | 89.9 | 30.7 |

TABLE V

mAP (%) RESULTS OF OUR METHOD USING DIFFERENT BACKBONE NETWORKS ON DUKEMTMC-REID

| Backbones | MobileNet | InceptionV2 | ResNet50 |
|---|---|---|---|
| Baseline | 54.8 | 61.7 | 66.6 |
| Ours | 68.2 | 73.0 | 77.1 |

attribute information, our method could achieve about 7% mAP improvement and about 3% R1 improvement on Market-1501 and DukeMTMC-reID datasets. This is because our proposed model with AAB can select key attributes and only leverage beneficial attribution information. Although CUHK03 dataset does not have attribute annotation, our method performs surprisingly well, which implies the potential of our method in discovering latent attributes in an unsupervised manner. Moreover, our method outperforms all the baselines and achieves the state-of-the-art results on all three datasets.

### *E. Analyses on Network Parameters*

In Table IV, we compare the number of parameters used in different models and their performance, including our proposed model and its variations: "w/o AAB. w/o PBM", "w/o AAB", "w/o ASM" and our full-fledged model, which show that our model does not introduce too many parameters and the number of parameters still remains in the same order of magnitude as backbone. For more comparisons, we investigate the inference time and the number of parameters for the baseline methods. We estimate the number of parameters of AANet [18], CA3Net [54] and APR [17] based on the papers. Their model scales are close to ours, but their performance on DukeMTMC-ReID dataset are much worse than ours, which shows the effectiveness and efficiency of our proposed model.

### *F. Effects of Different Backbone Networks*

In this subsection, we run experiments to prove that our proposed AAB can also be integrated into backbones like MobileNet [62] and InceptionV2 [63] other than ResNet on DukeMTMC-reID dataset. We are not applying PBM [34] or PCB [33] to remove the effects of different frameworks. The results are shown in Table V. "Baseline" is with the same setting for "Stage 1; w/o AAB, w/o PBM" and "Ours" is with the same setting for "Stage 1 + 2 + 3; w/o PBM" in Table II. The results show that our proposed method is effective for different backbones.

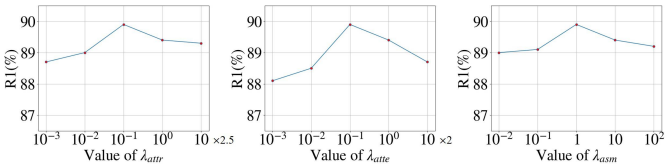


Fig. 5. The performance of our method by varying $\lambda_{attr}$, $\lambda_{atte}$, and $\lambda_{asm}$ on DukeMTMC-reID dataset.

### *G. Analyses for Hyperparameters*

To analyze different hyperparameters in Eqn. (16), we run experiments to investigate the impact of $\lambda_{attr}$, $\lambda_{atte}$ and $\lambda_{asm}$. We vary each hyperparameter and fix all the other ones to plot the change of R1 results. By taking DukeMTMC-reID dataset as an example, we report the results in Figure 5. R1 of our proposed method is in the range of [89.0, 89.9] when setting $\lambda_{attr}$ in the range of $2.5 \times [10^{-2}, 10]$. R1 is in the range of [88.5, 89.9] when setting $\lambda_{atte}$ in the range of $2 \times [10^{-2}, 10]$. R1 is in the range of [89.0, 89.9] when setting $\lambda_{asm}$ in the range of $[10^{-2}, 10^{2}]$. We can see that $\lambda_{atte} = 0.2$, $\lambda_{attr} = 0.25$, and $\lambda_{asm} = 1$ are the best choices and the performance variance is very small when setting these hyperparameters in a reasonable range.

### *H. Transferring Attribute Information Between Different Datasets*

Market-1501 and DukeMTMC-reID datasets are collected in totally different conditions (summer *v.s.* autumn). Some of the attributes are not transferable (*e.g.*, "Boots" and "Lower Length"). Also, the style for other attributes is also different due to different conditions (see Figure 8). We found that CUHK03 and Market-1501 are more similar to each other for the attributes appearances. To test whether the learned attribute classifier can be transferred to another dataset, we use a model trained on Market-1501 and DukeMTMC-reID to predict the pseudo attribute labels for CUHK03 respectively. Then we leverage these pseudo labels to train a new model on CUHK03 using our proposed approach. Since the numbers of attributes are different for Market1501 (12) and DukeMTMC-reID (10), we also vary the number of attributes we used for CUHK03. In Table VI, "12 Attributes; w/o transfer" and "10 Attributes; w/o transfer" are the results for our full-fledged method without attribute supervision (by setting $\lambda_{attr} = 0$ in Eqn. (16).). While "12 Attributes; Market1501 $\rightarrow$ CUHK03" and "10 Attributes; DukeMTMC-reID $\rightarrow$ CUHK03" are the results obtained by transferring attribute classifier from Market1501 and DukeMTMC-reID to CUHK03 respectively. We can see that without attribute supervision, the results of using 12 and 10 attributes are close, which means that the impact of using different numbers of unsupervised attributes is minor. However, the results of "Market1501 $\rightarrow$ CUHK03" is better than the others while "DukeMTMC-reID $\rightarrow$ CUHK03" is worse. Recall that the attribute appearances in CUHK03 are

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

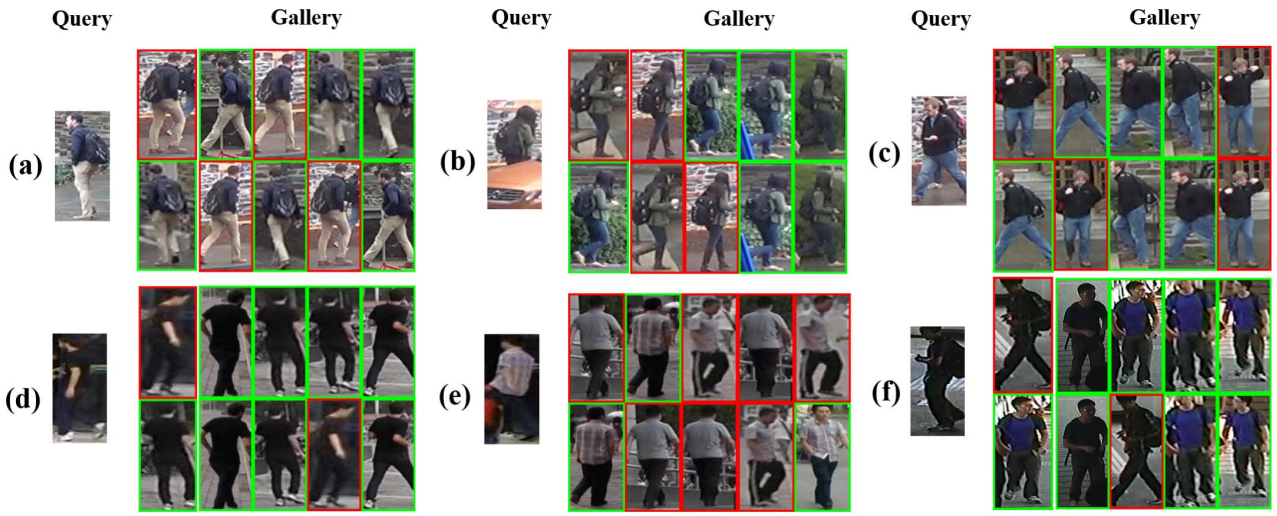


Fig. 6. Six Re-ID example queries. We show the top-5 results ranking from left to right. Images in green boxes are positive results while red boxes are negative results. The first three examples (a) (b) (c) are selected from DukeMTMC-reID dataset, (d) and (e) are selected from Market1501 dataset while (f) is selected from CUHK03 dataset. In each example, the upper parts are the retrieval results of our special case ("w/o AAB") and the lower parts are the retrieval results of our full-fledged method.

TABLE VI

RESULTS FOR ATTRIBUTE TRANSFER FROM MARKET1501 AND DUKEMTMC-REID TO CUHK03 RESPECTIVELY

| | 12 Attributes | | 10 Attributes | |
|---|---|---|---|---|
| | w/o transfer | Market1501 → CUHK03 | w/o transfer | DukeMTMC-reID → CUHK03 |
| MAP(%) | 78.1 | 79.0 | 78.0 | 77.5 |
| R1(%) | 80.1 | 81.0 | 80.2 | 79.8 |

similar to Market1501 but different from DukeMTMC-reID, we conclude that the attribute classifier can be transferred between datasets with similar attribute appearance (*e.g.*, Market1501 → CUHK03).

### *I. Qualitative Analyses*

Figure 6 shows six Re-ID visual examples collected from all the three datasets. The gallery images are split into two parts: the upper parts are the retrieval results by the baseline model ("w/o AAB"), while the lower parts are the retrieval results by our full-fledged model. The red (*resp.*, green) bounding boxes mean wrong (*resp.*, correct) retrievals. We can see that our model correctly ranks the positive images higher compared with the baseline model on both three datasets.

We visualize the tSNE results of baseline ResNet50 ("Stage 1; w/o AAB" in Table II) and our full-fledged model in Figure 7. The features we used is the image embedding $\mathbf{f}$ (see Section III-D). We choose 20 different person IDs from the test set of DukeMTMC-reID to visualize the feature distribution by tSNE. Different colors and numbers indicate different IDs. Our proposed method can better distinguish the identities like 7, 8, 16, 17, and 19, which are hard to be distinguished by the baseline method.

In Figure 8, we visualize the attention heat map for "FAB" ("AAB→FAB"), "AAB" ("Stage 1 Ours"), "AAB + ASM" ("Stage 1 + 2 + 3 Ours") by averaging and normalizing the attention map $\mathbf{A}$ in Eqn. (4). In each sample, we take the leftmost image as the input to visualize the attention heat map and list the noisy attributes dropped by ASM.

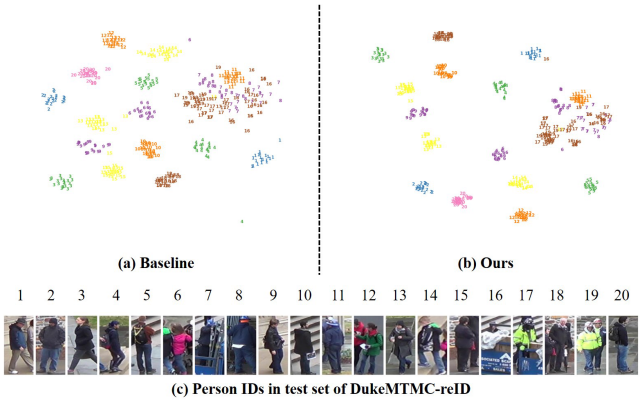


Fig. 7. tSNE visualization results for the baseline model and our proposed method on DukeMTMC-reID dataset.

In Figure 8(a), we show six examples from DukeMTMC-reID dataset. In particular, we show the attention map of top-1 retrieved results in Figure 6(a), (b) and (c). We can see that with the help of AAB and ASM, our model can distinguish persons with more attention to the key attributes (*e.g.*, shoe) (see left column), drop the noisy attributes which may vanish due to the occlusion by car (see middle column), or suppress the non-discriminative attributes as both persons wear black upper clothes and blue lower clothes (see right column).

In Figure 8(b), we show three examples from Market1501 dataset. Overall speaking, the attention heat maps show that the attention generated by AAB is more diverse than FAB. ASM will affect AAB by dropping noisy attributes to concentrate on more discriminative regions.

In Figure 8(c), we show three examples from CUHK03 dataset, which is not associated with attribute annotation. We visualize the attention map generated by our model without attribute supervision and report the number of dropped latent attributes. We observe that AAB can still generate diverse attention maps, but it is hard to provide a clear semantic explanation based on latent attributes.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

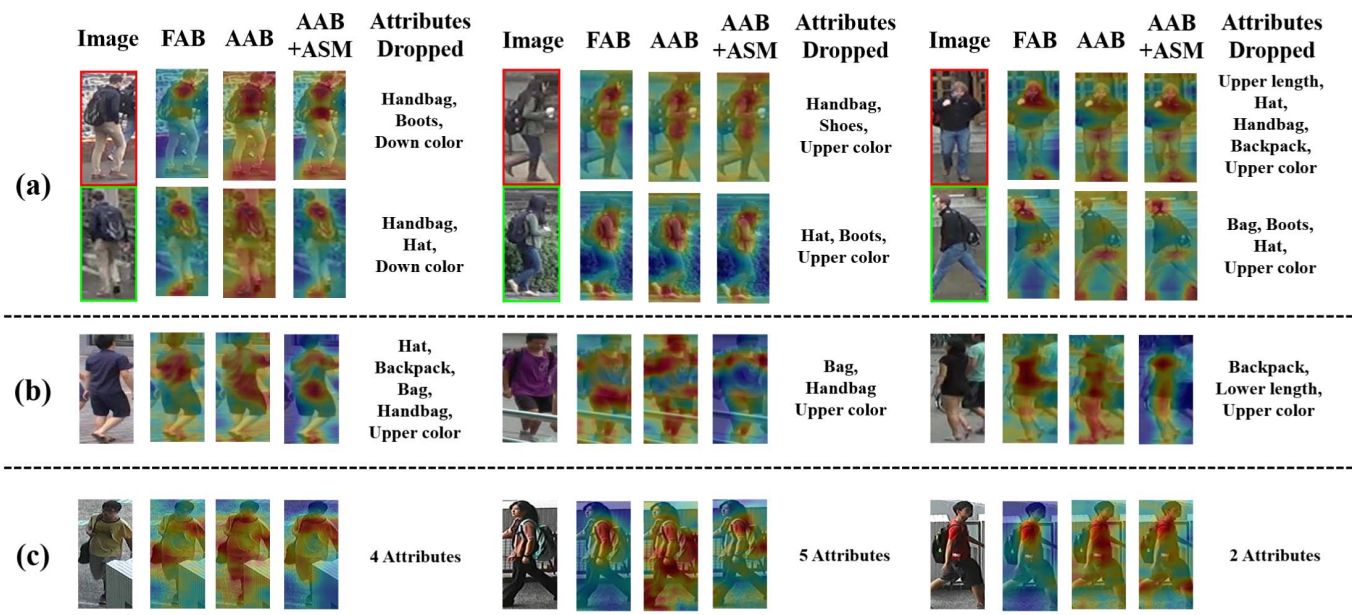


Fig. 8. Visualization of the attention heat map and the attributes dropped by ASM. The examples in (a), (b), and (c) are selected from DukeMTMC-reID, Market1501, and CUHK03 dataset respectively.

TABLE VII

ATTRIBUTE ANALYSES FOR THREE DATASETS. WE LIST THE ATTRIBUTE CLASSIFICATION RESULTS OF APR [17], AANET [18] AND OUR "STAGE1" MODEL (IN "AAB" COLUMN). ALSO, WE LIST THE PROBABILITY OF KEEPING THE CORRESPONDING ATTRIBUTES AFTER EACH EPISODE IN "ASM" COLUMN WHICH CAN REVEAL THE IMPORTANCE OF THE ATTRIBUTES IN RE-ID TASK

| Market-1501 | | | | | DukeMTMC-reID | | | | CUHK03 |
|---|---|---|---|---|---|---|---|---|---|
| Attribute name | Classification accuracy APR [17] (%) | AANet [18] (%) | AAB (%) | Keep ratio ASM (%) | Attribute name | Classification accuracy APR [17] (%) | AAB (%) | Keep ratio ASM (%) | Keep ratio ASM (%) |
| Gender | 88.9 | 92.3 | 93.1 | 94.2 | Gender | 84.2 | 86.2 | 95.2 | 76.3 |
| Hair length | 84.4 | 86.6 | 85.2 | 90.1 | Upper length | 88.4 | 89.2 | 33.9 | 72.1 |
| Sleeve length | 93.6 | 94.5 | 94.6 | 32.1 | Boots | 87.5 | 88.0 | 67.2 | 70.9 |
| Lower length | 93.7 | 94.2 | 92.2 | 29.0 | Hat | 87.6 | 90.2 | 33.1 | 65.2 |
| Lower type | 92.8 | 94.8 | 92.5 | 45.8 | Backpack | 75.8 | 80.2 | 56.2 | 60.1 |
| Hat | 97.1 | 98.0 | 98.2 | 3.2 | Bag | 82.9 | 81.3 | 20.7 | 55.1 |
| Backpack | 84.9 | 87.8 | 87.9 | 20.3 | Handbag | 93.4 | 95.5 | 3.5 | 52.8 |
| Bag | 76.4 | 79.7 | 79.0 | 17.7 | Shoes color | 89.7 | 90.2 | 63.1 | 51.9 |
| Handbag | 90.4 | 89.6 | 89.9 | 13.0 | Upper color | 74.2 | 73.2 | 40.9 | 45.2 |
| Age | 88.6 | 88.2 | 90.3 | 93.2 | Lower color | 69.9 | 71.6 | 38.7 | 44.0 |
| Upper color | 74.0 | 77.1 | 75.2 | 43.9 | - | - | - | - | - |
| Lower color | 73.8 | 70.8 | 74.4 | 52.4 | - | - | - | - | - |
| Average | 86.6 | 87.8 | 87.7 | 44.7 | Average | 83.4 | 85.6 | 45.3 | 59.4 |

Overall, compared with "FAB", "AAB" can pay more attention to the whole image to get diverse information, then our proposed model will select the most discriminative parts based on the attention map generated by AAB. Compared with "AAB", "AAB + ASM" pays less attention to the regions which are related to the dropped attributes. Combined with the dropped attributes shown in Figure 8 and the query results in Figure 6, we can see how ASM selects attributes and changes the attention map obtained by AAB. Also, we analyze the role of each attribute in Re-ID task, which will be demonstrated in the next section.

### *J. Analyses on Attribute Selection*

In Table VII, we densely analyze the attributes with the results of classification accuracy by APR [17], AANet [18] and our proposed AAB ("Stage 1 Ours"). APR and AANet both take attribute recognition as an auxiliary task for Re-ID task, while our proposed AAB model is also supervised with attribute information based on Equation (1) and can generate attention map at the same time. All the models are trained on the training set and evaluated on the test set for Market1501 and DukeMTMC-reID dataset for both Re-ID task and attribute recognition task with the annotations provided by [17]. For CUHK03 dataset, we train our model without attribute supervision and evaluate ASM on the test set. Notice that AAB only uses information from the first four stages in ResNet50 while APR and AANet use all the five stages information. As shown in Table VII, AAB can classify the attributes better than APR and competitive with AANet.

We also list the probability for each attribute not to be dropped in an episode in "ASM" column in Table VII. The attribute with higher keep ratio are more likely to help our proposed model in Re-ID task.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

We found that for some attributes like "Hat" in Market1501 and "Handbag" in DukeMTMC-reID, the probability is extremely low because the attributes are heavily imbalanced. Fewer than 5% of persons are wearing hats in Market1501 or carrying handbags in DukeMTMC-reID. These attributes are discriminative but rare. If query images have (*resp.*, don't have) these rare attributes, keeping these attributes will be helpful (*resp.*, unhelpful) to filter out the wrong gallery samples. However, most of the query images don't have these rare attributes. Focusing on these attributes will be unhelpful to most of the query images. So ASM tends to drop these rare attributes. While for the attributes containing rich semantic information, like "Gender" and "Age", our proposed model tends to keep them for better Re-ID performance. The attributes like "Gender" and "Age" are more balanced compared to the attributes like "Hat", and contains more discriminative information (*e.g.*, pose and hairstyle) compared to the attributes like "Lower color". Our method can provide insightful guidance for attribute annotation. For example, we can only annotate those balanced and semantically discriminative attributes.

## V. Conclusion

In this paper, we have proposed an Attribute Attentional Module (AAB) to enhance the performance of person Re-ID task leveraging identity-level attribute information, in which diverse attention maps are learned based on identity and identity-level attribute information. Moreover, noisy attributes are discarded by our designed Attribute Selection Module (ASM) sequentially using reinforcement learning. Comprehensive experiments on three benchmark datasets have demonstrated the superiority of our method.

## References


[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: https://arxiv.org/abs/2001.04193

[2] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[3] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[4] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 144–151.

[5] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Trans. Multimedia*, early access, doi: 10.1109/TMM.2020.2969782.

[6] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *J. Mach. Learn. Res.*, vol. 6, no. 1, pp. 937–965, 2006.

[7] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3610–3617.

[8] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[9] Z. Wang *et al.*, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 3006–3020, Oct. 2018.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114.

[11] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[13] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2285–2294.

[14] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. 15th Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 365–381.

[15] T. Chen *et al.*, "ABD-net: Attentive but diverse person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct. 2019, pp. 8351–8361.

[16] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 1389–1398.

[17] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[18] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 7134–7143.

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 2015, pp. 1116–1124.

[20] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis.*, Boston, MA, USA, Oct. 2017, pp. 3774–3782.

[21] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Image Anal. 17th Scandin. Conf.*, Ystad, Sweden, May 2011, pp. 91–102.

[22] S. Khamis, C. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," in *Proc. 13th Eur. Conf. Comput. Vis. Workshop*, Zurich, Switzerland, Sep. 2014, pp. 134–146.

[23] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2288–2295.

[24] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 152–159.

[25] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[26] W. Liao, M. Y. Yang, N. Zhan, and B. Rosenhahn, "Triplet-based deep similarity learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 385–393.

[27] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.

[28] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 1288–1296.

[29] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 3988–3994.

[30] Y. Chen, N. Wang, and Z. Zhang, "DarkRank: Accelerating deep metric learning via cross sample similarities transfer," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LO, USA, Feb. 2018, pp. 2852–2859.

[31] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 3980–3989.


Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.

[32] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 907–915.
[33] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 501–518.
[34] F. Zheng *et al.*, "Pyramidal person re-IDentification via multi-loss dynamic training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 8514–8522.
[35] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4806–4817, Oct. 2017.
[36] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[37] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.
[38] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 407–419, 2020.
[39] M. Ye and P. C. Yuen, "PurifyNet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2655–2666, 2020.
[40] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. 28th Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 2204–2212.
[41] X. Lan, H. Wang, S. Gong, and X. Zhu, "Deep reinforcement learning attention selection for person re-identification," 2017, *arXiv:1707.02785*. [Online]. Available: http://arxiv.org/abs/1707.02785
[42] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 2194–2200.
[43] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin, "Localizing by describing: Attribute-guided attention localization for fine-grained recognition," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 4190–4196.
[44] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 475–491.
[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. 29th IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2921–2929.
[46] T. Matsukawa and E. Suzuki, "Person re-identification using cnn features learned from combination of attributes," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancún, Mexico, Dec. 2016, pp. 2428–2433.
[47] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," 2015, *arXiv:1511.05952*. [Online]. Available: http://arxiv.org/abs/1511.05952
[48] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6781–6789.
[49] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 5735–5744.
[50] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.
[51] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
[52] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2011.
[53] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. 30th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 2094–2100.
[54] J. Liu, Z.-J. Zha, H. Xie, Z. Xiong, and Y. Zhang, "CA$_3$Net: Contextual-attentional attribute-appearance network for person re-identification," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, Seoul, Republic of Korea, Oct. 2018, pp. 737–745.
[55] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3652–3661.
[56] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896
[57] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[58] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2013–2025, 2020.
[59] Y. Zhang, Q. Zhong, L. Ma, D. Xie, and S. Pu, "Learning incremental triplet margin for person re-identification," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, Jan. 2019, pp. 9243–9250.
[60] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, Jan. 2019, pp. 8295–8302.
[61] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 2138–2147.
[62] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861
[63] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 448–456.

**Jianfu Zhang** received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 2015, where he is currently pursuing the Ph.D. degree with the Computer Science and Engineering Department. His current research interests include deep learning, adversarial learning, and generative models.

**Li Niu** received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2011, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2017. He was a Post-Doctoral Associate with Rice University, Houston, TX, USA. He is currently an Associate Professor with the Computer Science and Engineering Department, Shanghai Jiao Tong University, Shanghai, China. His current research interests include machine learning, deep learning, and computer vision.

**Liqing Zhang** (Member, IEEE) received the Ph.D. degree from Sun Yet-sen University, Guangzhou, China, in 1988. He was a Research Scientist with the RIKEN Brain Science Institute, Japan, from 1997 to 2002. He is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include visual cognitive computing, uncertainty reasoning, and statistical machine learning.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on December 21,2020 at 07:32:10 UTC from IEEE Xplore. Restrictions apply.