Pattern Recognition 95 (2019) 151–161

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Improving person re-identification by attribute and identity learning

Yutian Lin[a], Liang Zheng[b], Zhedong Zheng[a], Yu Wu[a], Zhilan Hu[a], Chenggang Yan[c], Yi Yang[a,*]

[a] Center for Artificial Intelligence, University of Technology Sydney, Australia
[b] Australian National University, Australia
[c] Hangzhou Dianzi University, China

ARTICLE INFO

Article history:
Received 15 May 2018
Revised 1 April 2019
Accepted 5 June 2019
Available online 6 June 2019

Keywords:
Person re-identification
Attribute recognition

ABSTRACT

Person re-identification (re-ID) and attribute recognition share a common target at learning pedestrian descriptions. Their difference consists in the granularity. Most existing re-ID methods only take identity labels of pedestrians into consideration. However, we find the attributes, containing detailed local descriptions, are beneficial in allowing the re-ID model to learn more discriminative feature representations. In this paper, based on the complementarity of attribute labels and ID labels, we propose an attribute-person recognition (APR) network, a multi-task network which learns a re-ID embedding and at the same time predicts pedestrian attributes. We manually annotate attribute labels for two large-scale re-ID datasets, and systematically investigate how person re-ID and attribute recognition benefit from each other. In addition, we re-weight the attribute predictions considering the dependencies and correlations among the attributes. The experimental results on two large-scale re-ID benchmarks demonstrate that by learning a more discriminative representation, APR achieves competitive re-ID performance compared with the state-of-the-art methods. We use APR to speed up the retrieval process by ten times with a minor accuracy drop of 2.92% on Market-1501. Besides, we also apply APR on the attribute recognition task and demonstrate improvement over the baselines.

1. Introduction

Person re-ID [1–4] and attribute recognition [5–7] both imply critical applications in surveillance. Person re-ID is a task of finding the queried person from non-overlapping cameras, while the goal of attribute recognition is to predict the presence of a set of attributes from an image. Attributes describe detail information for a person, including gender, accessory, the color of clothes, *etc.* Two examples of how attributes describe a person are shown in Fig. 1 (a). In this paper, we aim to improve the performance of large-scale person re-ID, using complementary cues from attribute labels. The motivation of this paper is that existing large-scale pedestrian datasets for re-ID contains only annotations of identity labels, we believe that attribute labels are complementary with identity labels in person re-ID.

The effectiveness of attribute labels is three-fold: First, training with attribute labels improves the discriminative ability of a re-ID model. The ID label can only coarsely define the distances among all the identities. This is not optimal since the appearance similarity of identities is overlooked. For example, as shown in Fig. 1(b), the bottom two pedestrians are very similar to each other, and they look very different from the top one. However, with only identity labels, the three pedestrians are uniformly distributed in the target space, which may harm model training. A more natural way is to treat these pedestrians differently according to their similarity. Attribute labels can depict pedestrian images with more detailed descriptions. These local descriptions push pedestrians with similar appearances closer to each other and those different away from each other (Fig. 1(c)). Second, detailed attribute labels explicitly guide the model to learn the person representation by designated human characteristics. With only identity labels and no detailed descriptions, the re-ID model have to infer the differences of pedestrians by itself, which is hard to learn a good semantic feature representation for persons. With the attribute labels, the model is able to learn to classify the pedestrians by explicitly focusing on some local semantic descriptions, which greatly ease the training of models. Third, attributes can be used to accelerate the retrieval process of re-ID. The main idea is to filter out some gallery images that do not have the same attributes as the query.

Several datasets are released for the pedestrian attribute. Li et al. [8] release a large-scale pedestrian attribute dataset RAP. Since the RAP dataset does not have ID labels, it is usually used

[*] Corresponding author.
E-mail address: yi.yang@uts.edu.au (Y. Yang).

https://doi.org/10.1016/j.patcog.2019.06.006
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

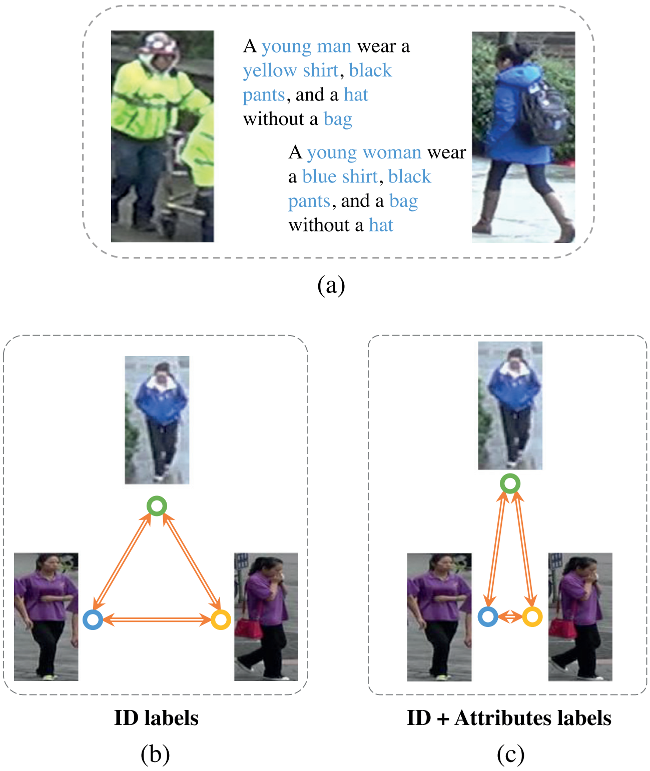

A young man wear a yellow shirt, black pants, and a hat without a bag
A young woman wear a blue shirt, black pants, and a bag without a hat
(a)
ID labels
(b)
ID + Attributes labels
(c)

Fig. 1. (a) Two examples of how attributes describe a person. (b) The three pedestrians are of different identities. Guided with only ID labels, images of three different identities have the same label distance between each other. (c) The three pedestrians are of different identities. Guided with ID and attribute labels, the bottom two identities are getting closer to each other in target space while the top one is pushed far away.

to transfer attribute knowledge to the target re-ID dataset. In [6], the PETA dataset is proposed which contains both attribute and identity attributes. However, PETA is comprised of small datasets and most of the datasets only contain one or two images for an identity. The lack of training images per identity limits the deep learning research. When using attributes for re-ID, attributes can be used as auxiliary information for low level features [9] or used to better match images from two cameras [10–12]. In recent years, some deep learning methods are proposed [13–15]. In these works, the network is usually trained by several stages. Franco et al. [13] propose a coarse-to-fine learning framework. The network is comprised of a set of hybrid deep networks, and one of the networks is trained to classify the gender of a person. In this work, the networks are trained separately and thus may overlook the complementarity of the general ID information and the attribute details. Besides, since gender is the only attribute used in the work, the correlation between attributes is not leveraged in [13]. In [14,15], the network is first trained on an independent attribute dataset, and then the learned information is transferred to the re-ID task. A work closest to ours consists of [16], in which the CNN embedding is only optimized by the attribute loss. We will show that by combining the identification and attribute recognition with an attribute re-weighting module, the APR network is superior to the method proposed in [16].

Comparing with previous methods, our paper differs in two main aspects. First, our work systematically investigates how person re-ID and attribute recognition benefit each other by a jointly learned network. On the one hand, identity labels provide global descriptions for person images, which have been proved effective for learning a good person representation in many re-ID works [17–19]. On the other hand, attribute labels provide detailed local descriptions. By exploiting both local (attribute) and global (identity) information, one is able to learn a better representation for a person, thereby achieving higher accuracy for person attribute recognition and person re-ID. Second, in previous works, the correlations of attributes are hardly considered. In fact, many attributes usually co-occur for a person, and the correlations of attributes may be helpful to re-weight the prediction of each attribute. For example, the attributes “skirt” and “handbag” are highly related to “female” rather than “male”. Given these gender-biased attribute descriptions, the probability of the attribute “female” should increase. We thereby introduce an Attribute Re-weighting Module to utilize correlations among attributes and optimize attribute predictions.

In this paper, we propose the attribute-person recognition (APR) network to exploit both identity labels and attribute annotations for person re-ID. By combining the attribute recognition task and identity classification task, the APR network is capable of learning more discriminative feature representations for pedestrians, including global and local descriptions. Specifically, we take attribute predictions as additional cues for the identity classification. Considering the dependencies among pedestrian attributes, we first re-weight the attribute predictions and then build identification upon these re-weighted attributes descriptions. The attribute is also used to speed up the retrieval process by filtering out the gallery images with different attribute from the query image. In the experiment, we show that by applying the attribute acceleration process, the evaluation time is saved to a significant extent. We evaluate the performance of the proposed method APR on two large-scale re-ID datasets and an attribute recognition dataset. The experimental results show that our method achieves competitive re-ID accuracy to the state-of-the-art methods. In addition, we demonstrate that the proposed APR yields improvement in the attribute recognition task over the baseline in all the testing datasets.

Comparing with existing works, our contributions are summarized as follows:

(1) We have manually labeled a set of pedestrian attributes for the Market-1501 dataset and the DukeMTMC-reID dataset. Attribute annotations of both datasets are publicly available on our website (https://vana77.github.io).
(2) We propose a novel attribute-person recognition (APR) framework. It learns a discriminative Convolutional Neural Network (CNN) embedding for both person re-identification and attributes recognition.
(3) We introduce the Attribute Re-weighting Module (ARM), which corrects predictions of attributes based on the learned dependency and correlation among attributes.
(4) We propose an attribute acceleration process to speed up the retrieval process by filtering out the gallery images with different attribute from the query image. The experiment shows that the size of the gallery is reduced by ten times, with only a slight accuracy drop of 2.92%.
(5) We achieve competitive accuracy compared with the state-of-the-art re-ID methods on two large-scale datasets, i.e., Market-1501 [17] and DukeMTMC_reID [20]. We also demonstrate improvements in the attribute recognition task.

2. Related work

CNN-based person re-ID. CNN-based methods are dominating the re-ID community upon the success of deep learning [20–26]. A branch of works learning deep metrics [27–29] that image pairs or triplets are fed into the network. Usually, the spatial constraints are integrated into the similarity learning process [28,30]. For example, in [23], a gating function is inserted in each convolutional layer, so that some subtle difference between two input images can be captured. Generally speaking, deep metric learning methods have advantages in training on relatively small datasets, but its efficiency on larger galleries may be compromised. Another branch of works learning deep representations [20,24,25,31]. Xiao et al. [24] propose to learn a generic feature embedding by training a classification model from multiple domains with a domain guided dropout. In [20], the combination of verification and classification losses is proven effective. Xu et al. [32] propose a Pose guided Part Attention (PPA)is learned to extract attention-aware feature for body parts from a base network. Then the features of body parts are further re-weighted, resulting in the final feature vector. Since GAN proposed by Goodfellow et al. [33], methods utilizing GAN [34,35] have been proposed to tackle re-ID. In [34], a Person Transfer Generative Adversarial Network (PTGAN) is proposed to transfer the image style from one dataset to another while keeping the identity information to bridge the domain gap. In [36], a dictionary-learning scheme is applied to transfer the feature learned by object recognition and person detection (source domains) to person re-ID (target domain). Recently, some semi-supervised methods [19,37] and unsupervised methods [22,38] has been proposed to address the data problem for re-ID. These methods achieve surprising performances with less or none of annotations. Attributes information also benefits these methods in the semi-supervised task.

In this paper, we adopt the simple classification model as our baseline and further exploit the mutual benefit between the traditional identity label and the attribute label.

Attributes for person re-ID. In some early attempts, attributes are used as auxiliary information to improve low-level features [11,12,39,40]. In [9,39], low-level descriptors and SVM are used to train attribute detectors, and the attributes are integrated by several metric learning methods. Su et al. [11,12] utilize both low-level features and camera correlations learned from attributes for re-identification in a systematic manner. In [41], a dictionary learning model is proposed that exploits the discriminative attributes for the classification task. Recently, some deep learning methods are proposed. Franco et al. [13] propose a coarse-to-fine learning framework, which is comprised of a set of hybrid deep networks. The network is trained for distinguishing person/not person, predicting the gender of a person and person re-ID, respectively. In this work, the networks are trained separately and might overlook the complementarity of the ID label and the attribute label. Besides, gender is the only attribute used in the work, so that the correlation between attributes is not leveraged. However, these works do not consider the correlation between attributes nor show if the proposed method improves the attribute recognition baselines. In [14], Su et al. first train a network on an independent dataset with attribute label, and then fine-tune the network the target dataset using only identity label with triplet loss. Finally, the predicts attribute labels for the target dataset is combined with the independent dataset for the final round of fine-tuning. Similarly, in [15], the network is pre-trained on an independent dataset labeled with attributes, and then fine-tuned on another set with person ID. In [42], a set of attribute labels are used as the query to retrieve the person image. Adversarial learning is used to generate image-analogous concepts for query attributes and get it matched with the image in both the global level and semantic ID level. The attribute is also used as supervision for unsupervised learning. Wang et al. [43] propose an unsupervised re-ID method that shares the source domain knowledge through attributes learned from labelled source data and transfers such knowledge to unlabelled target data by a joint attribute identity transfer learning across domains.

3. Attribute annotation

We manually annotate the Market-1501 [17] dataset and the DukeMTMC-reID [35] dataset with attribute labels. Although the Market-1501 and DukeMTMC-reID datasets are both collected on university campuses, and most identities are students, they are significantly different in seasons (summer vs. winter) and thus have distinct clothes. For instance, many people wear dresses or shorts in Market-1501, but most of the people wear pants in DukeMTMC-reID. So for the two datasets, we use two different sets of attributes. The attributes are carefully selected considering the characteristics of the datasets, so that the label distribution of an attribute (e.g., wearing a hat or not) is not heavily biased.

For Market-1501, we have labeled 27 attributes: Gender (male, female), hair length (long, short), sleeve length (long, short), length of lower-body clothing (long, short), type of lower-body clothing (pants, dress), wearing hat (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), carrying other types of bag (yes, no), 8 colors of upper-body clothing (black, white, red, purple, yellow, gray, blue, green), 9 colors of lower-body clothing (black, white, red, purple, yellow, gray, blue, green, brown) and age (child, teenager, adult, old). Positive and negative examples of some representative attributes of the Market-1501 dataset are shown in Fig. 2.

For DukeMTMC-reID, we have labeled 23 attributes: Gender (male, female), shoe type (boots, other shoes), wearing hat (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), carrying other types of bag (yes, no), color of shoes (dark, bright), length of upper-body clothing (long, short), 8 colors of upper-body clothing (black, white, red, purple, gray, blue, green, brown) and 7 colors of lower-body clothing (black, white, red, gray, blue, green, brown).

Note that all the attributes are annotated at the identity level. For example, in Fig. 2, the first two images in the second row are of the same identity. Although we cannot see the backpack clearly in the second image, we still annotate there is a "backpack" in the image. For both Market-1501 and DukeMTMC-reID, we illustrate the attribute distribution in Fig. 3. We define correlation of two attributes as the possibility that they co-occur on a person. We show the correlations between some representative attributes in Fig. 4. Attribute pairs with higher correlation are in a darker grid.

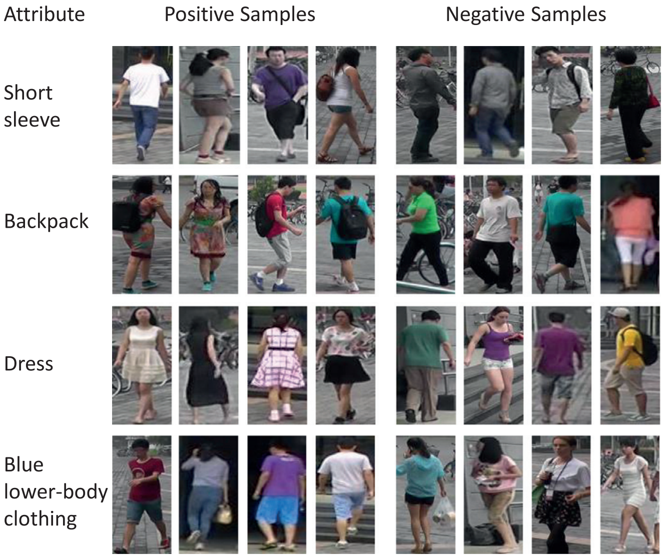
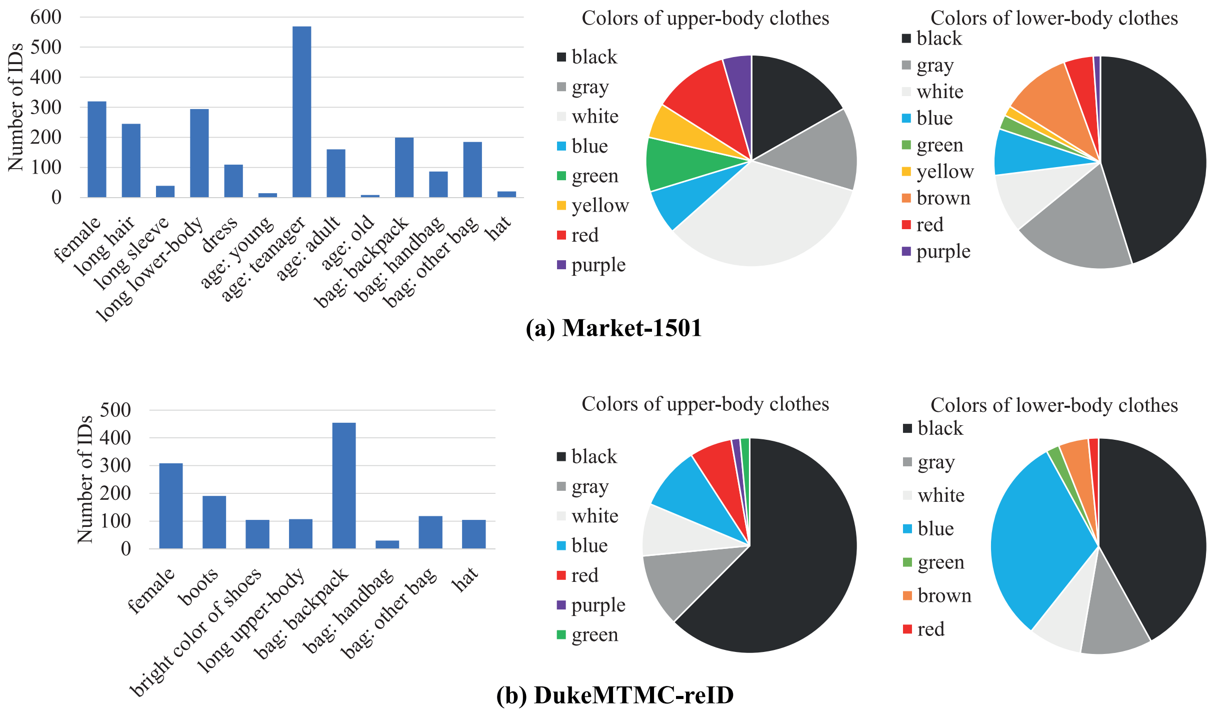


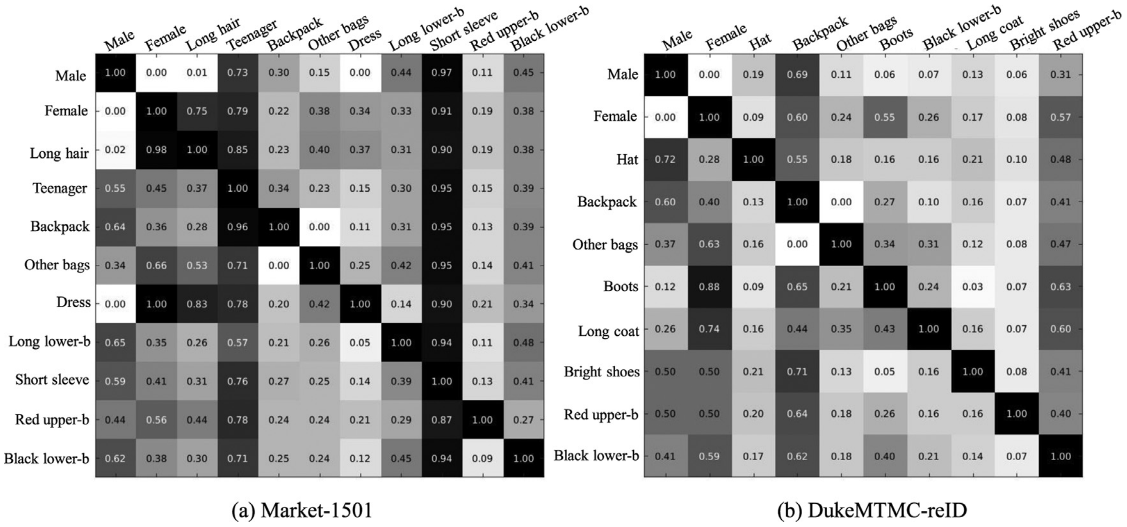
Fig. 2. Positive and negative examples of some representative attributes: *short sleeve, backpack, dress, blue lower-body clothing*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Market-1501

(b) DukeMTMC-reID

Fig. 3. The distributions of attributes on (a) Market-1501 and (b) DukeMTMC-reID. The left figure of each row shows the numbers of positive IDs for attributes except the color of upper/lower-body clothing. The middle and right pie chart illustrate the distribution of the colors of upper-body clothing and lower-body clothing, respectively.



(a) Market-1501

(b) DukeMTMC-reID

Fig. 4. Attribute correlations on the Market-1501 and DukeMTMC-reID datasets. A larger value indicates a higher correlation between the two attributes. We only show some of the representative attributes in the figure.

4. The proposed method

We first describe the necessary notations and two baseline methods in Section 4.1 and then introduce our proposed Attribute-Person Recognition network in Section 4.2. Finally, we introduce the attribute acceleration process in Section 4.3.

4.1. Preliminaries

Let $\mathcal{S}_{\mathcal{I}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the pedestrian identity labeled data set, where x_i and y_i denotes the i-th image and its identity label, respectively. For each image $x_i \in \mathcal{S}_{\mathcal{I}}$, we have the attributes annotations $\boldsymbol{a}_i = (a_i^1, a_i^2, \ldots, a_i^m)$, where a_i^j is the j-th attribute label for the image x_i, and m is the number of attributes classes. Let $\mathcal{S}_{\mathcal{A}} = \{(x_1, \boldsymbol{a}_1), \ldots, (x_n, \boldsymbol{a}_n)\}$ be the attribute labeled set. Note that set $\mathcal{S}_{\mathcal{I}}$ and set $\mathcal{S}_{\mathcal{A}}$ share common pedestrian images $\{x_i\}$. Based on these two set $\mathcal{S}_{\mathcal{I}}$ and $\mathcal{S}_{\mathcal{A}}$, we have the following two baselines:

Baseline 1 ID-discriminative Embedding (IDE). Following [17], we take IDE to train the re-ID model, which regards re-ID training process as an image identity classification task. It is trained only on the identity label data set $\mathcal{S}_{\mathcal{I}}$. We have the following objective function for IDE:

$$\min_{\boldsymbol{\theta}_I, \boldsymbol{w}_I} \sum_{i=1}^{n} \ell(f_I(\boldsymbol{w}_I; \phi(\boldsymbol{\theta}_I; x_i)), y_i), \tag{1}$$

where ϕ is the embedding function, parameterized by $\boldsymbol{\theta}_I$, to extract the feature from the data x_i. CNN models [17,20] are usually used as the embedding function ϕ. f_I is an identity classifier, parameterized by $\boldsymbol{w}_I$, to classify the embedded image feature $\phi(\boldsymbol{\theta}_I; x_i)$ into a k-dimension identity confidence estimation, in which k is the number of identities. ℓ denotes the suffered loss between classifier prediction and its ground truth label.

Baseline 2 Attribute Recognition Network (ARN). Similar to the IDE baseline for identity prediction, we propose the Attribute Recognition Network (ARN) for attribute prediction. ARN is trained only on the attribute label data set $\mathcal{S}_{\mathcal{A}}$. We define the following objective function for ARN:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}_A} \sum_{i=1}^{n}\sum_{j=1}^{m} \ell(f_{A_j}(\boldsymbol{w}_{A_j};\phi(\boldsymbol{\theta};x_i)),a_i^j), \qquad (2)$$

where f_{A_j} is the j-th attribute classifier, parameterized by $\boldsymbol{w}_{A_j}$, to classify the embedded image representation $\phi(\boldsymbol{\theta}; x_i)$ to the j-th attribute prediction. We take the sum of all the suffered losses for m attribute predictions on the input image x_i as the loss for the i-th sample.

In the evaluation stage of person re-ID task, for both baseline models, we use the embedding function $\phi(\boldsymbol{\theta}; \cdot)$ to embed the query and gallery images into the feature space. The query result is the ranking list of all gallery data according to the Euclidean Distance between the query data and each gallery data, i.e., $||\phi(\boldsymbol{\theta}; x_q) - \phi(\boldsymbol{\theta}; x_g)||_2$, where x_q and x_g denote the query image and the gallery image, respectively. For the evaluation of attribute recognition task, we take the attribute prediction $f_A(\boldsymbol{w}_A;\ \phi(\boldsymbol{\theta};\ \cdot))$ as the output, thereby evaluated with the ground truth by the classification metric.

4.2. Attribute-Person Recognition network

4.2.1. Architecture overview

The pipeline of the proposed APR network is shown in Fig. 5. APR network contains two prediction parts, one for attribute recognition task and the other for the identity classification task. Given an input pedestrian image, the APR network first extracts the person feature representation by the CNN extractor ϕ. Subsequently, APR predicts attributes based on the image feature. Here we calculate the attribute losses by the attribute prediction and ground truth labels. For the identity classification part, motivated by the fact that local descriptors (attributes) benefit global identification, we take the attribute predictions as additional cues for identity prediction. Specifically, to better leverage the attributes, given an input image, the APR network firstly computes attribute losses for the M individual attributes. Then the M prediction scores are concatenated and fed into an Attribute Re-weighting Module (ARM). The output of ARM is then concatenated with the global image feature for ID loss computation. The final identification is built upon the concatenated local-global feature.

4.2.2. Attribute re-weighting module

Suppose the set of attribute predictions for the image x is $\{\tilde{a}^1, \tilde{a}^2, \dots, \tilde{a}^m\}$, where $\tilde{a}^j \in [0, 1]$ is the j-th attribute prediction score from the attribute classifier f_{A_j}. We concatenate the prediction scores as vector $\tilde{\boldsymbol{a}}$, where $\tilde{\boldsymbol{a}} \in \mathbb{R}^{1\times m}$. Then the confidence score $\boldsymbol{c}$ for its prediction $\tilde{\boldsymbol{a}}$ is learned as,

$$\boldsymbol{c} = \mathtt{Sigmoid}(\boldsymbol{v}\tilde{\boldsymbol{a}}^T + \boldsymbol{b}), \qquad (3)$$

where $\boldsymbol{v} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{b} \in \mathbb{R}^{m\times 1}$ are trainable parameters, and the confidence score $\boldsymbol{c} \in \mathbb{R}^{m\times 1}$ is a set of learned weight. Therefore, the attribute re-weighting module transforms the original prediction $\tilde{\boldsymbol{a}}$ to a new prediction score as

$$\boldsymbol{a} = \boldsymbol{c} \circ \tilde{\boldsymbol{a}}^T, \qquad (4)$$

where $\circ$ is the element-wise multiplication. The re-weighted prediction score $\boldsymbol{a}$ is then concatenated with the global image representation for further identity classification.

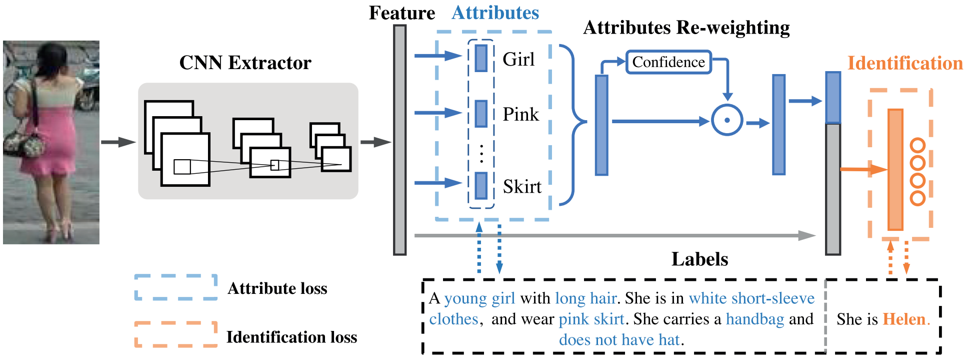
The motivation behind the Attribute Re-weighting Module (ARM) is to recalibrate the strengths of different activations of the attributes with a general consideration on all attributes. Therefore, we use trainable parameters ($\boldsymbol{v}$, $\boldsymbol{b}$) and the Sigmoid activation to perform a gating mechanism on the attribute predictions. With ARM, the model could learn to utilize the correlation between attributes. For instance, when the prediction scores of “pink upper-body clothes” and “long hair” are very high, the network may tend to up-weight the prediction scores for the attribute “female”.

4.2.3. Optimization

To exploit the attributes data $\mathcal{S}_{\mathcal{A}}$ as auxiliary annotations for the re-ID task, we propose Attribute-Person Recognition (APR) network. The APR network is trained on the combined data set $\mathcal{S}$ of the identity set $\mathcal{S}_{\mathcal{I}}$ and the attribute set $\mathcal{S}_{\mathcal{A}}$, i.e., $\mathcal{S} = \{(x_1, y_1, \boldsymbol{a}_1), \dots, (x_n, y_n, \boldsymbol{a}_n)\}$. For a pedestrian image x_i, we first extract the image feature representation by the embedding function $\phi(\boldsymbol{\theta};\ \cdot)$. Based on the image representation $\phi(\boldsymbol{\theta};\ x_i)$, two objective functions are optimized simultaneously:

The objective function for attribute predictions. Similar to the baseline ARN, the attribute predictions are obtained by a set of attribute classifiers on the input image feature, i.e., $\{f_{A_j}(\boldsymbol{w}_{A_j}; \phi(\boldsymbol{\theta}; x_i))\}$. We then optimize the objective function for attribute predictions the same as Eq. (2).

The objective function for identification. To introduce the attributes into identity prediction, we gather the attribute predictions $\{f_{A_j}(\boldsymbol{w}_{A_j}; \phi(\boldsymbol{\theta}; x_i))\}$ and re-weight them by the Attribute



A young girl with long hair. She is in white short-sleeve clothes, and wear pink skirt. She carries a handbag and does not have hat.

She is Helen.

Fig. 5. An overview of the APR network. APR contains two classification part, one for attribute recognition and the other for identification. Given an input image, the person feature representation is extracted by the CNN extractor ϕ. Subsequently, the attribute classifiers predict attributes based on the image feature. Here we calculate the attribute classification losses by the attribute predictions and ground truth labels. For the identity classification part, we take the attribute predictions as additional cues. Specifically, we first re-weight the local attribute predictions by the Attribute Re-weighting Module and then concatenate them with the global image feature. The final identification is built upon the concatenated local-global feature.

Re-weighting Module. We combine the re-weighted attribute predictions $\boldsymbol{a}_i$ and the image global feature $\phi(\boldsymbol{\theta}; x_i)$ to form a local-global representation. The identity classification is built upon the new feature. Thus we have the following objective function for identity prediction:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}_I} \sum_{i=1}^{n} \ell(\hat{f}_I(\hat{\boldsymbol{w}}_I; \hat{\boldsymbol{a}}_i, \phi(\boldsymbol{\theta}; x_i)), y_i), \tag{5}$$

where $\hat{\boldsymbol{a}}_i = (\hat{a}_i^1, \hat{a}_i^2, \dots, \hat{a}_i^m)$ is the concatenation of the re-weighted attribute predictions. $\hat{f}_I$ is the identity classier, parameterized by $\hat{\boldsymbol{w}}_I$, to predict the identity based on attribute predictions $\hat{\boldsymbol{a}}_i$ and image embeddings $\phi(\boldsymbol{\theta}; x_i)$.

The overall objective function. Considering both attribute recognition and identity prediction, we define the overall objective function as followings:

$$\min_{\boldsymbol{\theta},\boldsymbol{w}_I,\boldsymbol{w}_A} \lambda \sum_{i=1}^{n} \ell(\hat{f}_I(\hat{\boldsymbol{w}}_I; \hat{\boldsymbol{a}}_i, \phi(\boldsymbol{\theta}; x_i)), y_i)$$
$$+ (1-\lambda)\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} \ell(f_{A_j}(\boldsymbol{w}_{A_j}; \phi(\boldsymbol{\theta}; x_i)), a_i^j), \tag{6}$$

where λ is a hyper-parameter to balance the identity classification loss and the attribute recognition losses. We empirically discuss the effectiveness of λ in Section 5.3.

4.3. Attribute acceleration process

In the real-world application, calculating the distance for retrieval has become the main cost for a re-ID system, which is unaffordable. Attributes can be used to speed up the evaluation process by filtering the gallery data based on attribute predictions. The main idea is to filter out some gallery images that do not have the same attributes as the query.

During off-line computation, we apply feature extraction and attribute prediction for the gallery images. We take the attribute predictions with high confidence values as reliable ones for both query and gallery images. Then we remove those gallery candidates whose reliable attributes are different from the query. It is clear that the predicted attribute tends to be reliable as the prediction score gets higher. Specifically, we denote τ to be the threshold value. When the confidence score is higher than τ, the attribute is taken as a reliable one. When an attribute is reliable for both the query and gallery image, we check if the two images have the same prediction on that attribute. If not, this candidate image is removed from the gallery pool.

In real-life applications, this threshold is a trade-off between efficiency and accuracy. An *aggressive* choice is to set the threshold to a very small value (close to 0). It removes most of the candidates and maintains only a few candidates in on-line matching. This is suitable for the application where the retrieval speed is the main focus. A *conservative* option is to set the threshold to a large value (close to 1). It means we only remove a few candidates that are different in the very reliable attribute predictions from the query. In the empirical studies on Market-1501, we speedup the retrieval process by over ten times with a minor accuracy drop of 2.92% by setting the threshold to 0.7.

5. Experimental results

5.1. Datasets and evaluation protocol

We conduct experiments on two large-scale person re-ID datasets Market-1501 [17] and DukeMTMC-reID [35] and one attribute recognition dataset PETA [6].

The Market-1501 dataset contains 19,732 images for 751 identities for training and 13,328 images for 750 identities for testing. For each image, 27 attributes are annotated. To validate the hyper-parameter λ in Eq. (6), we use 651 identities in training set for training and the other 100 identities are used as the validation set to determine the value of parameter λ. We then use this hyper-parameter in the normal 751/750 split.

The DukeMTMC-reID dataset is a subset of the DukeMTMC dataset [44], which is divided into 16,522 training images for 702 identities and 19,889 test images for 702 identities. Each image is annotated with 23 labels as we described.

The PETA dataset is a large person attribute recognition dataset that annotated with 61 binary attributes and 4 multi-class attributes for 19,000 images. Following [6], 35 most important and interesting attributes are used in our experiments. Since most identities have a few training images, and some only have one training image, PETA is not an ideal testbed for re-ID deep learning research. In this paper, to evaluate our method on PETA, we re-split the dataset for the re-ID task. We use 17,100 images of 4981 identities for the experiment. In our new split, 9500 images of 4558 identities are used for training, 423 images are used for the query, and 7177 images are used for the gallery.

Evaluation metrics. For the person re-ID task, the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) are used for evaluation. In the experiments, we use the evaluation package publicly available in [17,35]. For the attribute recognition task, we test the classification accuracy for each attribute. The gallery images are used as the testing set. When testing the attribute prediction on Market-1501, we omit the distractor (background) and junks images, since they do not have attribute labels. We report the averaged accuracy of all these attribute predictions as the overall attribute prediction accuracy.

5.2. Implementation details

In the experiments, we adopt ResNet-50 [21] and CaffeNet [45] as the CNN backbone, respectively. The network is initialized by ImageNet [46] pre-trained models. Taking ResNet-50 for example, we append a 512-dim fully connected layer followed by Batch Normalization, a dropout layer with the drop rate of 0.5 and ReLU, after the âǣpool5âg layer. The 512-dim fully connected layer is then concatenated with the 27-dim (for Market-1501) attribute prediction score. The 539-dim (512+27) feature is used for identity classification. The experiment based on the CaffeNet is conducted similarly. Finally, the classification layer with *k* class nodes is used to predict the identity. For each attribute, we adopt a fully connected layer after the “pool5” layer as the classifier for attribute prediction. When evaluating the APR network for the re-ID task, we take the vertical concatenation of the embedded feature and the re-weighted attribute predictions as the final feature representation for each image.

Following Zheng et al. [20], we adopt a similar training strategy. Specifically, when using ResNet-50, we set the number of epochs to 60. The batch size is set to 32. Learning rate is initialized to 0.01 and changed to 0.001 in the last 20 epochs. For CaffeNet, the number of epochs is set to 110. For the first 100 epochs, the learning rate is 0.1 and changed to 0.01 in the last ten epochs. The batch size is set to 128. Randomly cropping and horizontal flipping are applied on the input images during training.

5.3. Evaluation of Person Re-ID task

5.3.1. Comparison with the state-of-the-art methods

The comparison with the state-of-the-art algorithms on Market-1501 and DukeMTMC-reID is shown in Tables 1 and 2, respectively. On Market-1501, we obtain **rank-1 = 87.04%, mAP**

Table 1
Comparison with state of the art on Market-1501. - the respective papers use hand-crafted features, * the respective papers use self-designed networks. “w/o ARM” denotes APR without the attribute re-weighting module. “w/o attri” denotes APR without the attribute recognition loss.

Methods	Publish	Backbone	Rank-1	Rank-5	Rank-10	mAP
MBC [47]	(AVSS2017)	*	45.56	67	76	26.11
SML [48]	(ECCV2016)	–	45.16	68.12	76	–
SL [49]	(CVPR2016)	–	51.9	–	–	26.35
Attri [16]	(ICPR2016)	AlexNet	58.84	–	–	33.04
S-CNN [23]	(ECCV2016)	*	65.88	–	–	39.55
2Stream [20]	(TOMM2017)	Res50	79.51	90.91	94.09	59.87
Cont-aware [50]	(CVPR2017)	*	80.31	–	–	57.53
Part-align [51]	(ICCV2017)	GoogLeNet	81.0	92.0	94.7	63.4
SVDNet [52]	(ICCV2017)	Res50	82.3	92.3	95.2	62.1
GAN [35]	(ICCV2017)	Res50	83.97	–	–	66.07
EBB[53]	(CVPR2018)	Inception	81.2	94.6	97.0	–
DSR[54]	(CVPR2018)	Res50	82.72	–	–	61.25
AACN[32]	(CVPR2018)	GoogLeNet	85.90	–	–	66.87
Baseline 1	–	CaffeNet	54.76	73.28	82.04	28.75
Baseline 1	–	Res50	80.16	92.03	94.98	57.82
Baseline 2	–	Res50	49.76	70.07	77.767	23.95
APR	–	CaffeNet	59.32	78.26	85.03	32.85
APR (w/o attri)	–	Res50	81.03	91.29	94.28	58.74
APR (w/o ARM)	–	Res50	85.71	94.32	96.46	66.59
APR	–	Res50	**87.04**	**95.10**	**96.42**	**66.89**

Table 2
Comparison with the state of the art on DukeMTMC-reID with ResNet-50. - the respective papers use hand-crafted feature. Rank-1 accuracy (%) and mAP (%) are shown. “w/o ARM” denotes APR without the Attribute Re-weighting Module.

Methods	Backbone	Rank-1	mAP
BoW+kissme [17]	–	25.13	12.17
LOMO+XQDA [55]	–	30.75	17.04
AttrCombine [16]	AlexNet	53.87	33.35
GAN [35]	Res50	67.68	47.13
SVDNet [52]	Res50	**76.7**	**56.8**
Baseline 1	Res50	64.22	43.50
Baseline 2	Res50	46.14	24.17
APR (w/o ARM)	Res50	73.56	54.79
APR	Res50	73.92	55.56

Table 3
Person reID performance on PETA with ResNet-50. Rank-1 accuracy (%) and mAP (%) are shown. “w/o ARM” denotes APR without the Attribute Re-weighting Module.

Methods	Backbone	Rank-1	Rank-5	Rank-10	mAP
Baseline 1	Res50	53.90	68.32	73.04	49.60
Baseline 2	Res50	43.30	60.08	70.23	39.52
APR (w/o ARM)	Res50	56.91	72.10	78.48	53.81
APR	Res50	**58.05**	**78.34**	**75.73**	**55.84**

= 66.89% by APR using the ResNet-50 model. We achieve the best rank-1 accuracy and mAP among the competing methods. On DukeMTMC-reID, our results are **rank-1 = 73.92% and mAP = 55.56%** by APR using ResNet-50. Our method is thus shown to compare favorably with the state-of-the-art methods.

5.3.2. Comparison with the baselines

Results on the three datasets are shown in Tables 1–3.

First, we observe that Baseline 2 (ARN) yields decent re-ID performance, e.g., a rank-1 accuracy of 49.76% using ResNet-50 on Market-1501. Note that Baseline 2 only utilizes attribute annotations without ID labels. This illustrates that attributes are capable of discriminating between different persons.

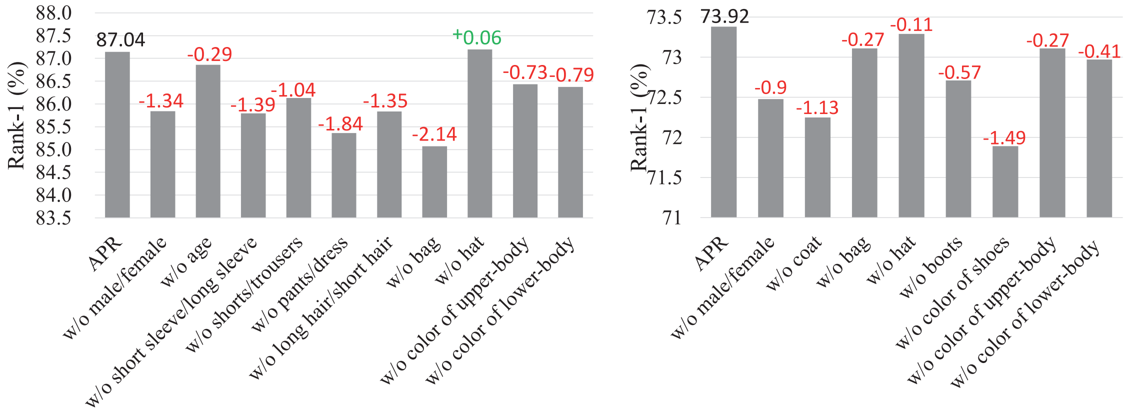
Second, by integrating the advantages in Baseline 1 and Baseline 2, our method exceed the two baselines by a large margin. For example, when using ResNet-50, the rank-1 improvement on Market-1501 over Baseline 1 and Baseline 2 is 6.88% and 37.28%, respectively. On DukeMTMC-reID, APR achieves 9.7% and 27.78% improvement over Baseline 1 and Baseline 2 in rank-1 accuracy. The consistent finding also holds for PETA, i.e., we observe improvements of 4.15% and 14.65% over Baseline 1 and Baseline 2 in rank-1 accuracy, respectively. This demonstrates the complementary nature of the two baselines, i.e., identity and attribute learning. We also observe that in Table 1, the performance of APR without attribute loss is slightly higher than that of B1. We believe that the slight improvement is lying on the difference of the network structure, that a Batch Normalization, a dropout layer and ReLU are further adopted in APR(w/o attri). However, the performance of both B1 and APR(w/o attri) still has a large margin between the performance of APR.

Third, for both backbone models (i.e., CaffeNet and ResNet-50), APR yields consistent improvement. On Market-1501, we obtain 4.56% and 6.88% improvements in rank-1 accuracy over Baseline 1 with CaffeNet and ResNet-50, respectively.

5.3.3. Ablation studies

Ablation study of attributes. We evaluate the contribution of individual attributes on the re-ID performance. We remove each attribute from the APR system at one time, and the results on the two datasets are summarized in Fig. 6. We find that most of the attributes on Market-1501 and DukeMTMC-reID are indispensable. The most influencing attribute on the two datasets are *bag types* and *the color of shoes*, which lead to a rank-1 decrease of 2.14% and 1.49% on the two datasets, respectively. This indicates that pedestrians of the two datasets have different appearances. The attribute of “wearing a hat or not” seems to exert a negative impact on the overall re-ID accuracy, but the impact is very small.

The effectiveness of the Attribute Re-weighting Module. We test APR with and without Attribute Re-weighting Module on the three re-ID datasets, and the results are shown in Tables 1–3. We observe performance improvement by using the Attribute Re-weighting Module for all the datasets. For Market-1501 with ResNet-50 as the backbone, the rank-1 and mAP improvements are 1.33% and 0.30%, respectively. For DukeMTMC-reID, the improvements are 0.36% and 0.74%, respectively. For PETA, we observe improvements of 1.14% and 2.03% in rank-1 and mAP, respectively. The improvement is consistent on all experiments.



(a) Market-1501

(b) DukeMTMC-reID

Fig. 6. Re-ID rank-1 accuracy on Market-1501 and DukeMTMC-reID. We remove one attribute from the system at a time. All the colors of upper-body clothing are viewed as one attribute here; the same goes for colors of lower-body clothing. Accuracy changes are indicated above the bars.

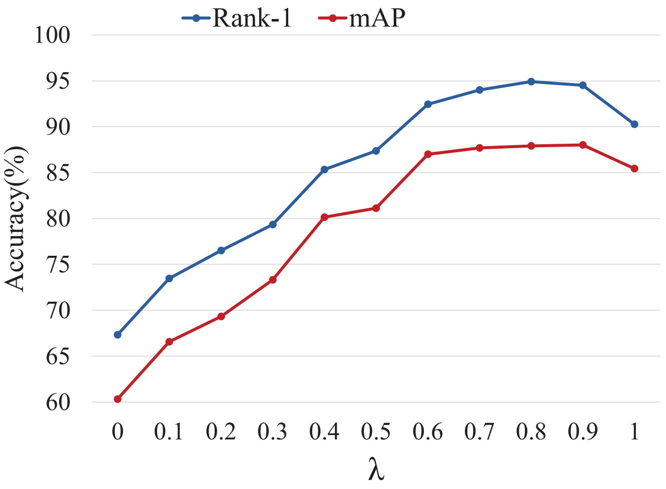


Fig. 7. The re-ID performance (rank-1 accuracy and mAP) curves on the validation set of Market-1501 with different values of parameter λ in Eq. (6). According to the performance curves, we set $\lambda = 0.9$ for all the experiment on Market-1501, DukeMTMC-reID and PETA.

5.3.4. Algorithm analysis

Parameter validation. We validate the parameter λ of APR on the validation set of Market-1501. λ is a key parameter balancing the contribution of the identification loss and attribute recognition loss (Eq. (6)). When λ becomes larger, person identity classification will play a more important role. Re-ID accuracy on the validation set of Market-1501 with different values of the parameter λ is presented in Fig. 7. We observe that when λ changes from 0 to 0.9, the rank-1 accuracy and mAP gradually increase from 67.33% and 60.32% to 94.52% and 88.03%, respectively. It indicates the importance of identity label in the re-ID task. When λ increases to 1, the rank-1 accuracy and mAP of the model decrease to 90.25% and 85.44% respectively, which indicates the effectiveness of attributes. The best re-ID performance is obtained when $\lambda = 0.9$. Therefore, we use $\lambda = 0.9$ for APR in all the following experiments.

Robustness of the learned representation in the Wild. To validate whether the proposed method still works under practical conditions, we report results on the Market-1501+500k dataset. The 500k distractor dataset is composed of background images and a large number of irrelevant pedestrians. The re-ID accuracy of our APR model with ResNet-50 on this dataset is presented in Fig. 8. It can be expected that the re-ID accuracy drops as the gallery gets larger due to more distractors. The results further show that our method outperforms both [20] and Baseline 1. However, the rank-1 accuracy of the proposed method drops faster than that of Baseline1. We think that the Baseline 1 may be able to retrieve the ground truths of easy queries, but APR could retrieve the ground truths of both the easy and hard queries. When increasing the number of images in the gallery, the easy query images can still be handled by both of the baseline and APR. However, the hard query sample can be harder to retrieve. Thus, the performance of APR drops faster.

5.3.5. Accelerating the retrieval process.

Fig. 9 illustrates the re-ID performance under different percentages of remaining gallery data. The number of remaining gallery images is controlled by the threshold τ. It helps indicate if an attribute is reliable. As τ increases, attributes with higher confidence score are taken as reliable ones to wipe out gallery images, and the number of remaining gallery images increases. As the percentage of remaining gallery data decreases from 78% to 8.7%, the rank-1 accuracy for re-ID decrease very slowly. When we try a more aggressive speedup, the performance drops quickly. For example, we observe an accuracy drop of 21.79% when we use only 0.5% gallery images. Note that with the remaining 8.68% gallery data, we still achieve 84.12% on rank-1 accuracy, which is close to the original result 87.04%.

In practice, most of the time is spent on calculating the distances between the query feature and the features of remained gallery images. In Market-1501, for instance, there are 3368 queries and 19,732 gallery images. Without the acceleration, the testing process takes 919.86 s (0.273 s per query), using an Intel i7-6850K CPU. With acceleration, there are only about 2000 images remaining in the gallery for each query, and it costs only 90.26 s (0.026 s per query) for testing. Although the saved time may be slight in the academic dataset, in real-world applications which involves a large amount of data, efficiency could be an important advantage.

5.4. Evaluation of attribute recognition

We test attribute recognition on the galleries of the Market-1501, DukeMTMC-reID, PETA in Tables 4–6, respectively. We also evaluate our method on the CUB_200_2011 dataset, which contains 11,788 images of 200 bird classes. Each category is annotated with 312 attributes, which are divided into 28 groups and are used as 28 multi-class attributes in our experiments. The result is shown in Table 7. By comparing the results of APR and Baseline 2 (ARN), two conclusions can be drawn:

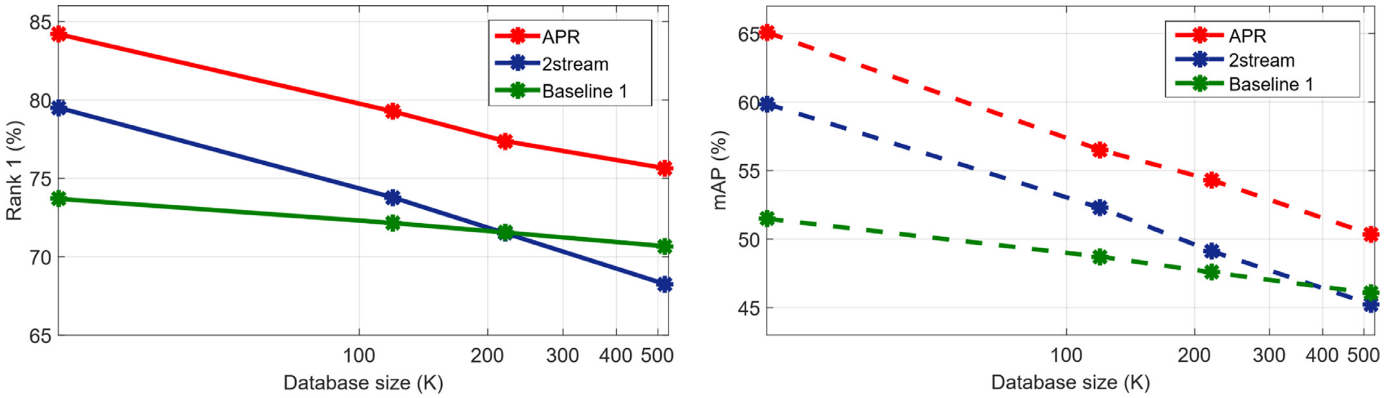


Fig. 8. Re-ID accuracy on the Market-1501+500k dataset. (Left:) rank-1 accuracy. (Right:) mean average precision. We compare our method with 2stream [20] and Baseline 1. As the number of images in the database increases, the accuracy of the three methods declines. However, APR remains the best performance.

Table 4
Attribute recognition accuracy on Market-1501. In "APR", parameter λ is optimized in Fig. 7. "L.slv", "L.low", "S.clth", "B.pack", "H.bag", "C.up", "C.low" denote *length of sleeve, length of lower-body clothing, style of clothing, backpack, handbag, color of upper-body clothing* and *color of lower-body clothing, resp*. "B2" denotes Baseline 2 (ARN).

	gender	age	hair	L.slv	L.low	S.clth	B.pack	H.bag	bag	hat	C.up	C.low	Avg
B2	87.5	85.8	84.2	93.5	93.6	93.6	86.6	88.1	78.6	97.0	72.4	71.7	86.0
APR	88.9	88.6	84.4	93.6	93.7	92.8	84.9	90.4	76.4	97.1	74.0	73.8	86.6

Table 5
Attribute recognition accuracy on DukeMTMC-reID. "L.up", "B.pack", "H.bag", "C.shoes", "C.up", "C.low" denote *length of sleeve, backpack, handbag, color of shoes, color of upper-body clothing* and *color of lower-body clothing, resp*. "B2" denotes Baseline 2 (ARN).

	Gender	Hat	Boots	L.up	B.pack	H.bag	Bag	C.shoes	C.up	C.low	Avg
B2	82.0	85.5	88.3	86.2	77.5	92.3	82.2	87.6	73.4	68.3	82.3
APR	84.2	87.6	87.5	88.4	75.8	93.4	82.9	89.7	74.2	69.9	83.4

First, on all datasets, the overall attribute recognition accuracy is improved by the proposed APR network to some extent. The improvements are 0.26%, 0.08%, 0.2% and 1.58% on Market-1501, DukeMTMC-reID, PETA and CUB_200_2011, respectively. So overall speaking, the integration of identity classification introduces some degree of complementary information and helps in learning a more discriminative attribute model. Also, note that we achieve the best attribute recognition result on PETA among the state-of-the-art.

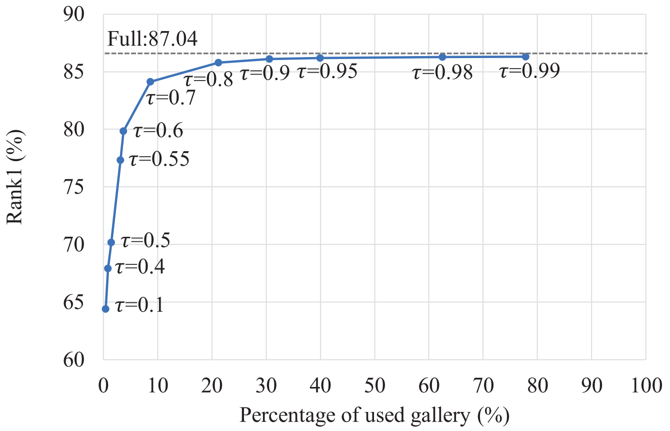


Fig. 9. Re-ID rank-1 accuracy curve on Market-1501 when using attributes to accelerate the retrieval process. For a query, we only take the gallery data with the same reliable attributes into consideration. The X-axis stands for different percentages of the *remaining* gallery data when using different filtering threshold values. Note that APR could speed up the retrieval process by nearly ten times (only 8.68% gallery data remains) with only a slight accuracy drop of 2.92%.

Table 6
Attribute recognition accuracy on PETA.

Attribute	Mean accuracy
MRFr2 [6]	71.1
ACN [56]	81.15
MVA [15]	84.61
Baseline 2	84.45
APR	**84.94**

Table 7
Attribute recognition accuracy on CUB_200_2011.

Methods	mean accuracy
Baseline 2	87.31
APR	**89.12**

Second, we observe that the recognition rate of some attributes decreases for APR, such as *hair* and *B.pack* in Market-1501. However, Fig. 6 demonstrates that these attributes are necessary for improving re-ID performance. The reason probably lies in the multi-task nature of APR. Since the model is optimized for re-ID (Fig. 7), ambiguous images of certain attributes may be incorrectly predicted. Nevertheless, the improvement on the two datasets is still encouraging and further investigations should be critical.

6. Conclusions and future work

In this paper, we mainly discuss how re-ID is improved by the integration of attribute learning. Based on the complementary of attribute labels and ID labels, we propose an attribute-person recognition (APR) network, which learns a re-ID embedding and predicts the pedestrian attributes under the same framework. We systematically investigate how the person re-ID and attribute recognition benefit each other. In addition, we re-weight the attribute predictions considering the dependencies and correlations among attributes of a person. To show the effectiveness of our method, we have annotated attribute labels on two large-scale re-ID datasets. The experimental results on two large-scale re-ID benchmarks demonstrate that by learning a more discriminative representation, APR achieves competitive re-ID performance compared with the state-of-the-art methods. We additionally use APR to accelerate the retrieval process of re-ID more than three times with a minor accuracy drop of 1.26% on Market-1501. For attribute recognition, we also observe an overall precision improvement using APR.

Pedestrian attributes provide a different view of the person re-identification problem. As a mid-level feature, attributes are more robust to environment changes, such as the background and illumination. In the future, we will first investigate the transferability and scalability of pedestrian attributes. For instance, we could adapt the attribute model learned on Market-1501 to other pedestrian datasets. Second, attributes provide a bridge to the image-text understanding. We will investigate a system using attributes to retrieve the relevant pedestrian images. It is useful in solving specific re-ID problems, in which the query image is missing and can be described by attributes.

References

[1] X. Zhu, B. Wu, D. Huang, W.-S. Zheng, Fast open-world person re-identification, IEEE Trans. Image Process. 27 (5) (2018) 2286–2300.

[2] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification, IEEE Trans. Image Process. 26 (7) (2017) 3492–3506.

[3] L. Wu, C. Shen, A. van den Hengel, Deep linear discriminant analysis on fisher networks: a hybrid architecture for person re-identification, Pattern Recognit. 65 (2017) 238–250.

[4] L. Ren, J. Lu, J. Feng, J. Zhou, Multi-modal uniform deep learning for RGB-D person re-identification, Pattern Recognit. 72 (2017) 446–457.

[5] J. Zhu, S. Liao, Z. Lei, S.Z. Li, Multi-label convolutional neural network based pedestrian attribute classification, Image Vis. Comput. 58 (2017) 224–229.

[6] Y. Deng, P. Luo, C.C. Loy, X. Tang, Pedestrian attribute recognition at far distance, in: Proceedings of the ACM international conference on Multimedia, 2014, pp. 789–792.

[7] A.H. Abdulnabi, G. Wang, J. Lu, K. Jia, Multi-task CNN model for attribute prediction, IEEE Trans. Multim. 17 (11) (2015) 1949–1959.

[8] D. Li, Z. Zhang, X. Chen, H. Ling, K. Huang, A richly annotated dataset for pedestrian attribute recognition, (2016) arXiv:1603.07054.

[9] R. Layne, T.M. Hospedales, S. Gong, Re-id: hunting attributes in the wild, in: The British Machine Vision Conference, 2014.

[10] S. Khamis, C.-H. Kuo, V.K. Singh, V.D. Shet, L.S. Davis, Joint learning for attribute-consistent person re-identification, in: European Conference on Computer Vision, 2014, pp. 134–146.

[11] C. Su, S. Zhang, F. Yang, G. Zhang, Q. Tian, W. Gao, L.S. Davis, Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping, Pattern Recognit. 66 (2017) 4–15.

[12] C. Su, F. Yang, S. Zhang, Q. Tian, L.S. Davis, W. Gao, Multi-task learning with low rank attribute embedding for multi-camera person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1167–1181.

[13] A. Franco, L. Oliveira, Convolutional covariance features: conception, integration and performance in person re-identification, Pattern Recognit. 61 (2017) 593–609.

[14] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Multi-type attributes driven multi–camera person re-identification, Pattern Recognit. 75 (2018) 77–89.

[15] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: The IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1435–1443.

[16] T. Matsukawa, E. Suzuki, Person re-identification using CNN features learned from combination of attributes, in: The IEEE International Conference on Pattern Recognition, 2016, pp. 2428–2433.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: The IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[18] S.-Z. Chen, C.-C. Guo, J.-H. Lai, Deep ranking for person re-identification via joint representation learning, IEEE Trans. Image Process. 25 (5) (2016) 2353–2367.

[19] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[20] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned CNN embedding for person reidentification, ACM Trans. Multim. Comput.Commun. Appl. 14 (1) (2017) 13.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[22] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: clustering and fine-tuning, ACM Trans. Multim. Comput. Commun. Appl. TOMCCAP 14 (4) (2018), doi:10.1145/3243316.

[23] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision, 2016, pp. 791–808.

[24] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.

[25] S. Zhou, J. Wang, D. Meng, X. Xin, Y. Li, Y. Gong, N. Zheng, Deep self-paced learning for person re-identification, Pattern Recognit. 76 (2018) 739–751.

[26] L. Zhu, Z. Xu, Y. Yang, A.G. Hauptmann, Uncovering the temporal context for video question answering, Int. J. Comput. Vis. 124 (3) (2017) 409–421.

[27] L. Ma, X. Yang, D. Tao, Person re-identification over camera networks using multi-task distance metric learning, IEEE Trans. Image Process. 23 (8) (2014) 3656–3670.

[28] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[29] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognit. 48 (10) (2015) 2993–3003.

[30] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.

[31] L. Wu, Y. Wang, J. Gao, X. Li, Deep adaptive feature embedding with local sample distributions for person re-identification, Pattern Recognit. 73 (2018) 275–288.

[32] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, Attention-aware compositional network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[34] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[35] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: The IEEE International Conference on Computer Vision, 2017, pp. 3774–3782.

[36] N. Martinel, M. Dunnhofer, G.L. Foresti, C. Micheloni, Person re-identification via unsupervised transfer of learned visual representations, in: Proceedings of the 11th International Conference on Distributed Smart Cameras, 2017, pp. 151–156.

[37] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, Y. Yang, Progressive learning for person re-identification with one example, IEEE Trans. Image Process. (2019). 1–1. doi: 10.1109/TIP.2019.2891895.

[38] Y. Lin, X. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: AAAI Conference on Artificial Intelligence, 2019.

[39] R. Layne, T.M. Hospedales, S. Gong, Q. Mary, Person re-identification by attributes, in: The British Machine Vision Conference, 2, 2012, p. 8.

[40] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, J. Bu, Attribute-restricted latent topic model for person re-identification, Pattern Recognit. 45 (12) (2012) 4204–4213.

[41] P. Peng, Y. Tian, T. Xiang, Y. Wang, T. Huang, Joint learning of semantic and latent attributes, in: European Conference on Computer Vision, 2016, pp. 336–353.

[42] Z. Yin, W.-S. Zheng, A. Wu, H.-X. Yu, H. Wan, X. Guo, F. Huang, J. Lai, Adversarial attribute-image person re-identification, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 1100–1106.

[43] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[44] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision, 2016, pp. 17–35.

[45] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 675–678.

[46] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[47] E. Ustinova, Y. Ganin, V. Lempitsky, Multi-region bilinear convolutional neural networks for person re-identification, in: The IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017, pp. 1–6.

[48] C. Jose, F. Fleuret, Scalable metric learning via weighted approximate rank component analysis, in: European Conference on Computer Vision, 2016, pp. 875–890.

[49] D. Chen, Z. Yuan, B. Chen, N. Zheng, Similarity learning with spatial constraints for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1268–1277.
[50] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7398–7407.
[51] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: The IEEE International Conference on Computer Vision, 2017, pp. 3239–3248.
[52] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: The IEEE International Conference on Computer Vision, 2017, pp. 3820–3828.
[53] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, X. Wang, Eliminating background-bias for robust person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[54] L. He, J. Liang, H. Li, Z. Sun, Deep spatial feature reconstruction for partial person re-identification: alignment-free approach, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[55] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.
[56] P. Sudowe, H. Spitzer, B. Leibe, Person attribute recognition with a jointly–trained holistic CNN model, in: The IEEE International Conference on Computer Vision Workshops, 2015, pp. 87–95.

Yutian Lin received the B.E. degree from Zhejiang University, China, in 2016. She is now a Ph.D. student in the Center for Artificial Intelligence, University of Technology Sydney, Australia. Her research interests are person re-identification and deep learning.

Liang Zheng received the Ph.D. degree in Electronic Engineering from Tsinghua University, China, in 2015 and the B.E. degree in Life Science from Tsinghua University China, in 2010. He was a postdoc researcher at the University of Texas at San Antonio, USA. He is now a postdoc researcher in the Center for Artificial Intelligence, University of Technology Sydney, Australia. His research interests are image retrieval, person re-identification and deep learning.

Zhedong Zheng is a Ph.D. student in the Center for Artificial Intelligence, University of Technology Sydney, Australia. His research interests are person re-identification and deep learning.

Yu Wu is a Ph.D. student in the Center for Artificial Intelligence, University of Technology Sydney, Australia. His research interests are video analysis and deep learning.

Zhilan Hu is a visiting researcher at the University of Technology Sydney, Australia. She received the Ph.D. degree from Tsinghua University in 2009. Her research is about person re-identification, face recognition and landmark detection.

Chenggang Yan received the B.S. degree in computer science from Shandong University in 2008 and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2013. He was an Assistant Research Fellow with Tsinghua University. He is currently a Professor with Hangzhou Dianzi Univeristy. He has authored or co-authored over 30 refereed journal and conference papers. His research interests include machine learning, image processing, computational biology, and computational photography.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a professor with the University of Technology Sydney, Australia. He was a post-doctoral researcher in the School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video content understanding.