Improving Person Re-identification by Attribute and Identity Learning

Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu and Yi Yang
University of Technology Sydney

`yutianlin477,liangzheng06,zdzheng12,wu08yu,yee.i.yang@gmail.com`

Abstract

Person re-identification (re-ID) and attribute recognition share a common target at the pedestrian description. Their difference consists in the granularity. Attribute recognition focuses on local aspects of a person while person re-ID usually extracts global representations. Considering their similarity and difference, this paper proposes a very simple convolutional neural network (CNN) that learns a re-ID embedding and predicts the pedestrian attributes simultaneously. This multi-task method integrates an ID classification loss and a number of attribute classification losses, and back-propagates the weighted sum of the individual losses.

Albeit simple, we demonstrate on two pedestrian benchmarks that by learning a more discriminative representation, our method significantly improves the re-ID baseline and is scalable on large galleries. We report competitive re-ID performance compared with the state-of-the-art methods on the two datasets.

1. Introduction

This paper aims to improve the performance of large-scale person re-identification (re-ID), using complementary cues from attribute labels . Both person re-ID and attribute recognition imply critical applications in surveillance. Person re-ID is a task of finding the queried person from the non-overlapping cameras, and the goal of attribute recognition is to predict the presence of a set of attributes from an image.

The major starting point of this paper is that person re-ID, especially those based on the CNN features, relies on global descriptors, while attribute recognition usually denotes local structures of a person. We speculate that correctly predicting person attributes can improve the discriminative ability of a re-ID system. A re-ID algorithm may fail to tell the subtle difference between two identities when their appearances look alike, but one can make a more precise judgment by looking into the details. As shown in the 4th row of Fig. 1, a re-ID system fails to discriminate between persons wearing similar blue and black clothes; but

Figure 1. Attribute improves re-ID. In the 3rd row, the two persons are featured by two distinct sets of attributes. In the 4th row, when a re-ID system fails to discriminate persons with similar appearance, attributes can offer complementary local information.

attributes may suggest that *male*, *not wearing a hat* and *no bags* can eliminate the false matches.

Comparing with the previous literature discussing re-ID and attributes, this paper differs in two aspects. First, most methods use attributes to strengthen the relationship of image pairs or triplets [33, 34, 16, 21]. Historically, this line of methods is such designed because the datasets usually provided only two images per identity. Yet the recent large-scale datasets (*e.g.*, Market-1501 [51] and DukeMTMC-reID [54]) provides richer training samples per class, and it is observed that training a classification model is superior to the siamese model [52]. Therefore, this paper adopts a classification CNN model to train the multi-task network. Second, to our knowledge, there are few works demonstrating the impact of the usage re-ID labels on attribute recognition, which also has critical research and application values. Our work makes an initial effort on whether re-ID can improve the accuracy of attribute recognition. Note that this paper mainly discusses ID-level attributes instead of instance-level attributes. ID-level attributes refer to those related to the person himself, such as gender, age, *etc*. Instance-level
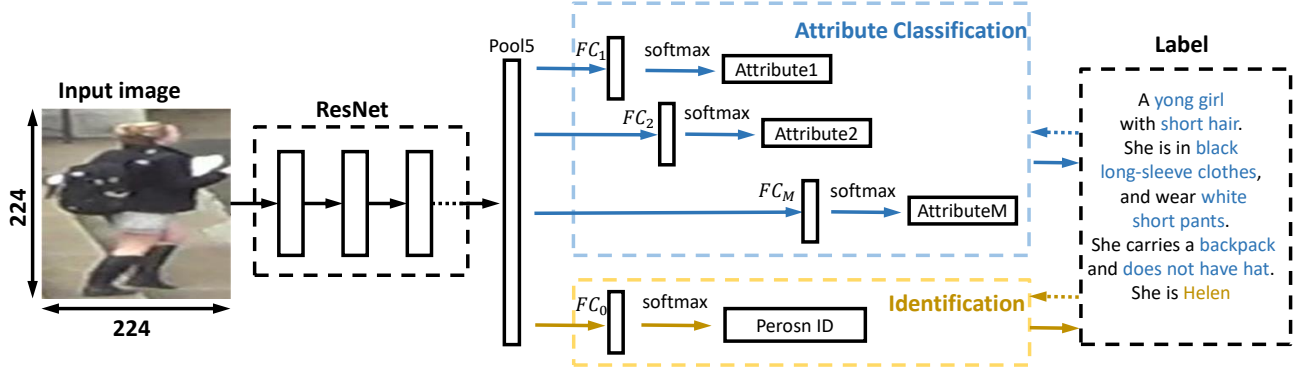
Figure 2. An overview of the APR network. During training, it predicts M attribute labels and an ID label. The weighted sum of the individual losses is back propagated. During testing, we extract the Pool5 (ResNet-50) or FC7 (CaffeNet) descriptors for retrieval.

attributes, in contrast, are those appearing for a short time or belonging to the external environment, *e.g.*, making a phone call, riding a bicycle. To some extent, person re-ID is a more generic task that takes attribute recognition into consideration, especially the ID-level attributes. In this sense, if two bounding boxes are of the same identity, we usually expect that most of the ID-level attributes should be matched. This thus may exert a positive effect on the recognition accuracy of most attributes.

In this paper, we offer a different view from previous works by mainly discussing how attribute labels help person re-ID in large-scale learning problems without resorting to image pairs. To our knowledge, this is the first work integrating attributes in the classification CNN model for re-ID. We propose the attribute-person recognition (APR) network which combines the two tasks on the loss level. The APR network is built upon two baselines, one for person re-ID, and the other for attribute recognition. Both baselines are implemented by a classification CNN architecture, and the re-ID baseline has been proven to yield competitive accuracy [52, 9, 2]. The APR network combines the person re-ID loss and attribute prediction loss (Fig. 2), so that their complementary aspects are leveraged to improve the re-ID accuracy. To evaluate the performance of the proposed method, we conduct experiments on the Market-1501 [51] and DukeMTMC-reID [54] datasets. We show that the learned embedding achieves competitive re-ID accuracy to the state-of-the-art methods. In addition, we demonstrate that the proposed APR network also demonstrates improvement in the attribute recognition performance over the baseline.

The main contributions are summarized as follows.

(1) Combining the ID and attribute classification losses, we propose a new attribute-person recognition (APR) network. It simultaneously learns a discriminative CNN embedding for re-ID and an attribute classification model, yielding competitive accuracy in re-ID and demonstrating

some improvement in attribute recognition.

(2) We have manually labeled a set of pedestrian attributes for the Market1501 and DukeMTMC-reID dataset. The attribute annotations will be made public available.

2. Related Work

This section briefly reviews several closely related aspects, *i.e.*, CNN-based re-ID methods, attributes for re-ID and attributes for face applications.

CNN-based person re-ID. CNN-based methods are dominating the re-ID community, and can be classified into two categories: deep metric learning and deep representation learning. For the first category, usually image pairs or triplets are fed into the network. Representative methods include [44, 23]. Usually, the spatial constraints are integrated into the similarity learning process [1, 23, 44, 5]. For example, in [38], a gating function is inserted in each convolutional layer, so that some subtle difference between two input images can be captured. In [5], Chen *et al*. propose a multi-task method by implementing a ranking loss and a verification loss from a triplet input. Generally speaking, deep metric learning methods have advantages in training on relatively small datasets, but its efficiency on larger galleries may be compromised.

The second category, *i.e.*, representation learning, has gained increasing popularity because it yields superior accuracy [52] and does not harm the efficiency. Examples include [41, 49, 42, 9, 53]. Xiao *et al*. [41] propose to learn a generic feature embedding by training a classification model from multiple domains with a domain guided dropout. In [53, 9], the combination of verification and classification losses is proven effective, consistent with the findings in [35]. This paper adopts this line of methods as the re-ID baseline, *i.e.*, a classification model is fine-tuned, and the learned embedding is used to compute the similarity between the query and gallery images.

Attributes for person re-ID. In person re-ID, attributes

have been investigated in a number of works. In most of them, attributes are used as auxiliary information to re-ID. In [21, 20, 19], low-level descriptors and SVM is used to train attribute detectors, and the attributes are integrated in several metric learning methods. Su *et al*. [33] learn a discriminative model by multi-task learning, which exploits features and attributes shared by different cameras. Khamis *et al*. [16] propose to jointly optimize the triplet loss for re-ID and the attribute classification loss, but it is not shown if the proposed method improves the attribute recognition baselines. These methods usually use image pairs or triplets for training, while our method employs the classification CNN model and analyzes the impact of re-ID on attribute recognition. Several datasets also released for these tasks. Deng *et al*. [7] and Li *et al*. [22] have released two large-scale pedestrian attribute datasets PETA and RAP. The PETA dataset does not contain an adequate number of training samples per ID, and RAP does not have ID labels, so we do not use the two datasets in this paper. Recently, Li [32] contribute a dataset composed of person images described by natural language. We do not use this dataset because we focus on the attribute recognition, but the natural language does not explicitly have clean attribute annotations.

The work closest to this paper consists in [36], in which the CNN embedding is learned only by the attribute loss. We will show that by simultaneously combining the ID and attribute classification losses, the APR network is superior to the method proposed [36].

Attributes for face applications. Attributes for face recognition have been studied for long. In the old days, Moghaddam *et al*. [29] propose to use the Haar features to predict gender by SVM. Lanitis *et al*. [18] compare various classifiers for age prediction. Recently, many deep learning methods have been proposed. Zhang *et al*. [48] use facial attribute recognition as an auxiliary task to improve the face alignment performance using the convolutional neural network. In [27], two CNN structures are cascaded and fine-tuned jointly with attribute tags to predict face attributes. Yang *et al*. [43] train CNNs for facial attribute recognition in order to obtain high responses in regions of faces, so that candidate windows of faces can be localized. But due to the complex CNN structure, this approach is time costly in practice.

3. Attribute Annotation

We manually annotate the Market-1501 [51] and DukeMTMC-reID [54] datasets with attribute labels due to two reasons. First, the current largest pedestrian attribute dataset, RAP [22], does not contain ID labels. Second, the PETA dataset [7] is an ensemble of relatively small re-ID datasets such as VIPeR [12] and iLIDS [28]. For PETA, the number of training samples per ID is very limited, which
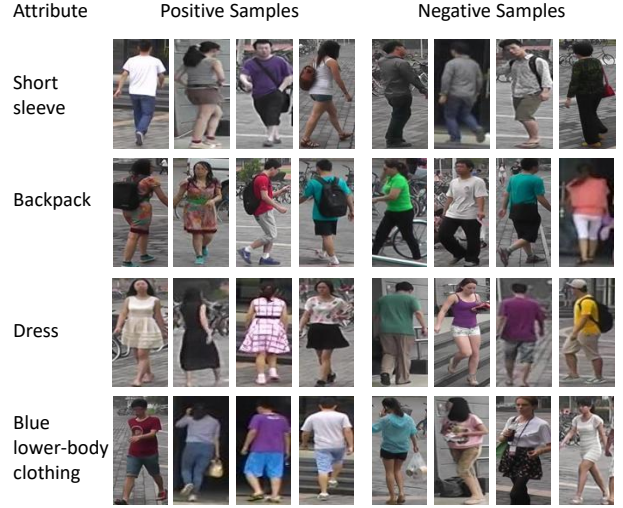

Figure 3. Positive and negative examples of some representative attributes: *short sleeve*, *backpack*, *dress*, *blue lower-body clothing*.

compromises the effectiveness of deep learning.

Although the Market-1501 and DukeMTMC-reID datasets collected in university campuses and most identities are students, they are significantly different in seasons (summer vs. winter) and thus have distinct clothes. For instance, many persons wear dresses or pants 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 bag (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), 8 colors of upper-body clothing (black, white, red, purple, yellow, gray, blue, green), 9 colors of lower-body clothing (black, white, pink, purple, yellow, gray, blue, green, brown) and age (child, teenager, adult, old). Note that the color attributes are binary ones. Positive and negative examples of some representative attributes of the Market-1501 dataset are shown in Fig. 3.

For DukeMTMC-reID, we have labeled 23 attributes: gender (male, female), shoe type (boots, other shoes), wearing hat (yes, no), carrying bag (yes, no), carrying backpack (yes, no), carrying handbag (yes, no), color of shoes (dark, light), 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). The color attributes are binary attributes too. For both Market-1501 and
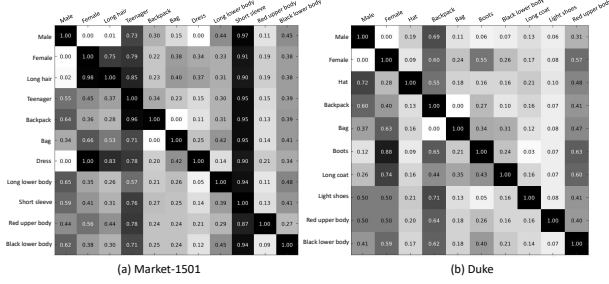
Figure 4. Attribute correlations on the Market-1501 and DukeMTMC-reID datasets. A larger value indicates higher correlation between the two attributes. Representative attributes are shown.

DukeMTMC-reID, we illustrate the correlations between some representative attributes in Fig. 4, and the attribute distribution of the two database are shown in Fig. 5.

Note that all the attributes are annotated in the identity level. For example, in Fig. 3, the first two images in the second row are of the same identity. Although we cannot see the backpack clearly in the second image, and the label of the image is still "backpack". Both the Market-1501 and DukeMTMC-reID attributes annotation are available on our website[1]

4. Proposed Method

We first describe two baselines in Section 4.1 and then the APR network in Section 4.2

4.1. Baseline Methods

This paper constructs two baselines for person re-ID and pedestrain attribute recognition. We use ResNet-50 [13] as the base network, as it is shown to yield competitive re-ID performance in [52]. The base network is pre-trained on ImageNet [6]. We fine-tune the two baselines using the newly annotated attributes and the currently available identity labels, respectively.

Baseline 1 (person re-ID). Given a base model, we set the number of neurons in the last fully-connected (FC) layer to K, where K denotes the number of training identities. To avoid overfitting, we insert a dropout layer before the FC layer, and set the dropout rate to 0.9. During testing, for each query and gallery image, we extract a 2,048-dim feature vector from the pool5 layer. For each query, we calculate the Euclidean distance between the query and gallery features, before a ranking step. The result of Baseline 1 is shown in Table 1.

Baseline 2 (pedestrian attribute recognition & re-ID). We use M FC layers followed by the softmax layers for

---

[1]https://vana77.github.io

attribute recognition, where M denotes the number of attributes. For CaffeNet, the M FC layers replace FC8. For ResNet-50, they replace the FC layer. For attributes with m classes, the FC layer is m-dim. We also insert a dropout layer as in Baseline 1. The result of Baseline 2 is shown in Table 3.

4.2. Attribute-Person Recognition (APR) Network

Architecture. In this section, we describe the proposed attribute-person recognition (APR) network. The APR network consists of a base model, $M + 1$ FC layers before loss computation, a loss for identity classification, and M losses for attribute classification, where M is the number of attributes. The new FC layers are denoted as FC_0, FC_1, ..., FC_M, where FC_0 is used for ID classification, and FC_1, ..., FC_M are used for attribute recognition. The dimensions of the new FC layers are the same with those in Baseline 1 and Baseline 2. Given an input image, the proposed network simultaneously predicts its identity and a set of attributes. The pre-trained model can be ResNet-50 [13] or CaffeNet [17].

For ResNet-50, as shown in Fig. 2, the FC layers are connected to Pool5. For CaffeNet, the FC layers are connected to FC7 instead. Images of size 224×224 and 227×227 are used for ResNet-50 and CaffeNet, respectively.

Loss computation. Suppose we have n images of K identities. Each identity has M attributes. Let $D_i = \{x_i, d_i, l_i\}$ be the training set where x_i denotes the i-th image, d_i denotes the identity of image x_i, and $l_i = \{l_i^1, ..., l_i^M\}$ is a set of M attribute labels of image x_i (as well as identity d_i).

Given a training example x, our model first computes its pool5 descriptor f (We take ResNet-50 as an example). The size of the outputted vector is $1 \times 1 \times 2048$. The output of the FC_0 layer is $\boldsymbol{z} = [z_1, z_2, ..., z_K] \in R^K$. So the predicted probability of each ID label $k \in 1, ..., K$ is calculated as: $p(k|x) = \frac{exp(z_k)}{\sum_{i=1}^{K} exp(z_i)}$. For brevity, let us omit the correlation between k and x. So the cross entropy loss of ID classification can be formulated as below:

$$L_{\text{ID}}(f, d) = -\sum_{k=1}^{K} log(p(k))q(k). \quad (1)$$

Let y be the ground-truth ID label, so that $q(y) = 1$ and $q(k) = 0$ for all $k \neq y$. In this case, minimizing the cross entropy loss is equivalent to maximizing the possibility of being assigned to the ground-truth class.

We also use M softmax losses for attribute prediction. We assume m classes for a certain attribute, and the probability of assigning sample x to the attribute class $j \in 1, ..., m$ can be written as $p(j|x) = \frac{exp(z_j)}{\sum_{i=1}^{m} exp(z_i)}$. Similarly,
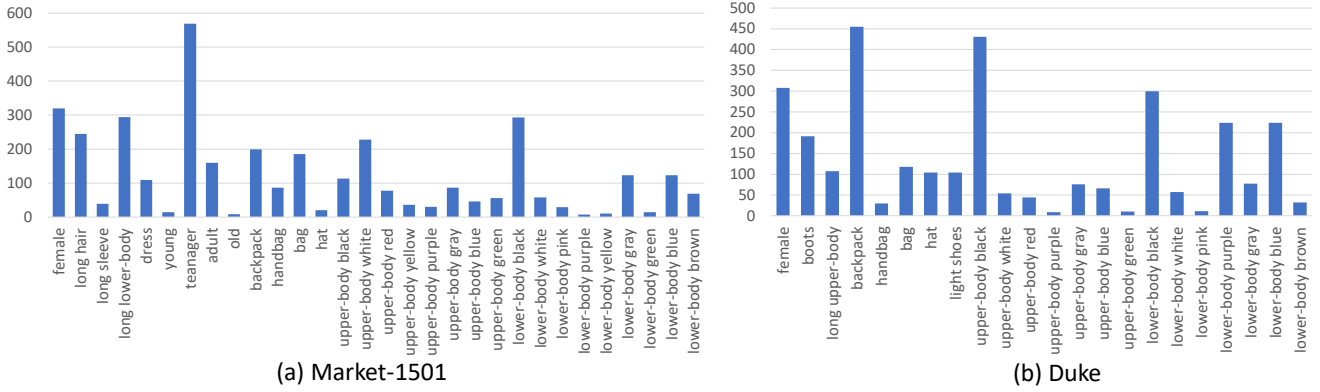
4

(a) Market-1501

(b) Duke

Figure 5. The distribution of attributes on (a) Market-1501 and (b) DukeMTMC-reID. For each attribute, we show the number of positive IDs.

dress   hat   backpack   short sleeve

Figure 6. Intermediate features maps learned in our network correspond to certain attributes.

the loss of classify sample x can be computed as below:

$$L_{\text{att}}(f, l) = -\sum_{j=1}^{m} log(p(j))q(j), \qquad (2)$$

Let y_m be the ground-truth attribute label, so that $q(y_m) = 1$ and $q(j) = 0$ for all $j \neq y_m$. The other symbols are the same as Eq. 1.

By using a multi-attribute classification loss function and an identity classification loss function, the APR network is trained to predict attribute and identity labels. Here the final loss function is defined as:

$$L = \lambda L_{\text{ID}} + \frac{1}{M}\sum_{i=1}^{M} L_{\text{att}}, \qquad (3)$$

Where L_{ID} and L_{att} denote the cross entropy loss of identity classification and attribute classification, respectively. Parameter λ balances the contribution of the two losses and is determined on a validation set of Market-1501.

We visualize the intermediate feature maps from CNN in Fig. 6, which shed light on how the integration of attribute enhances the interpretability of the network.

5. Experiment

5.1. Datasets and Evaluation Protocol

The Market1501 dataset [51], one of the largest person re-ID datasets, contains 32,668 gallery images and 3,368 query images captured by 6 cameras. It also includes 500k irrelevant images forming a distractor set, which may exert a considerable influence on the recognition accuracy. Market-1501 is split into 751 identities for training and 750 identities for testing. For most of our experiments, 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. When validating the re-ID performance, we randomly select one query image for each ID under each camera, so in total 431 queries are used in validation. We perform a cross-camera retrieval in both testing and validation.

The DukeMTMC-reID dataset [54] is a subset of the DukeMTMC dataset [31]. It contains 1,812 identities captured by 8 cameras. A number of 1,404 identities appear in more than two cameras, and the rest 408 IDs are distractor images. Using the evaluation protocol specified in [54], the training and testing sets both have 702 IDs. So in together, there are 2,228 query images, 16,522 training images and 17,661 gallery images.

Evaluation metrics. For the person re-ID task, we use the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP). For each query, its average precision (AP) is computed from its precision-recall curve. Then mAP is the mean value of average precisions across all queries. The presumption is that CMC reflects retrieval precision, while MAP reflects the recall. We use the evaluation package publicly available in [51, 54].

For the attribute recognition task, we test the classification accuracy for each attribute (24 and 21 attributes
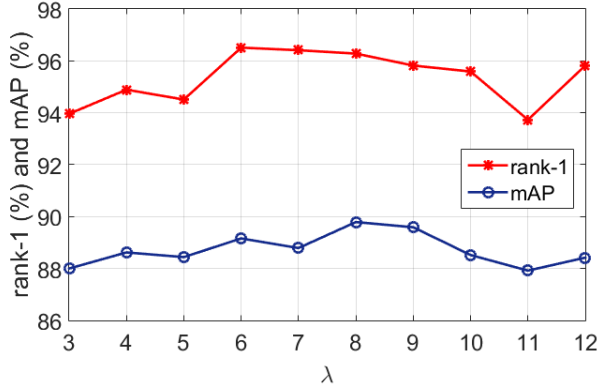
Figure 7. The re-ID accuracy (rank-1 accuracy and mAP) on the validation set of Market-1501 when parameter λ (Eq. 3) varies. We set $\lambda = 8$ on both Market-1501 and DukeMTMC-reID.

for Market-1501 and DukeMTMC-reID, respectively). The gallery images are used as the testing set. For Market-1501, the distractor (background) images and junks do not have attribute labels, so are not used for testing attribute prediction. We report the average recognition rate of all these attributes as the overall attribute prediction accuracy.

5.2. Implementation Details

We adopt the similar training strategy with [53]. Specifically, when using ResNet-50, we set the number of epochs to 55. The batch size is set to 64. Learning rate is initialized to 0.001 and changed to 0.0001 in the last 5 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 10 epochs. The batch size is set to 128. For both networks, the stochastic gradient descent (SGD) is implemented in each mini-batch to update the parameters.

5.3. Evaluation of Person Re-ID

Parameter validation. We first show the re-ID validation results of parameter λ which is a key parameter balancing the contribution of re-ID and attribute recognition (Eq. 3). When $\lambda = 0$, the APR network reduces to Baseline 2. When λ becomes larger, person identity classification will exerts more influence, and thus can approximate Baseline 1. Re-ID results on the validation set of Market-1501 is presented in Fig. 7. From the mAP and rank-1 results, we observe that both curves increase first, and then decrease. When $\lambda = 8$, a relatively higher re-ID performance can be obtained. Therefore, we use $\lambda = 8$ in both Section 5.3 and Section 5.4 if not specified.

Attribute recognition improves re-ID over the baselines. We evaluate if the APR network outperforms the two baselines. Results on the two datasets are shown in Table 1 and Table 2. Here we note that the FC descriptor of B2 can be used for re-ID the same way as the FC descriptor of B1.

Methods	rank-1	rank-5	rank-10	rank-20	mAP
DADM[34]	39.4	-	-	-	19.6
MBC[37]	45.56	67	76	82	26.11
SML[15]	45.16	68.12	76	84	-
DLDA[40]	48.15	-	-	-	29.94
SL[4]	51.9	-	-	-	26.35
DNS[45]	55.43	-	-	-	29.87
LSTM[39]	61.6	-	-	-	35.3
S-CNN[38]	65.88	-	-	-	39.55
2Stream[53]*	79.51	90.91	94.09	96.23	59.87
GAN[54]*	79.33	-	-	-	55.95
Pose[50]*	78.06	90.76	94.41	96.52	56.23
Deep[9]*	83.7	-	-	-	**65.5**
B1 (C, 651)	52.13	73.33	80.84	86.90	27.29
B1 (R, 651)	70.51	86.40	90.82	93.91	48.19
B1 (R, 751)	73.69	88.15	91.80	94.83	51.48
B2 (R, 651)	49.76	70.07	77.76	83.87	23.95
APR (C, 651)	57.54	78.26	85.03	90.38	32.85
APR (R, 651)	82.98	92.81	95.30	96.94	61.98
APR (R, 751)	**84.29**	**93.20**	**95.19**	**97.00**	64.67

Table 1. Comparison with state of the art on Market-1501. "B1" and "B2" denote Baseline 1 and Baseline 2, *resp*. "C" and "R" represent CaffeNet and ResNet-50, *resp*. The numbers in the bracket are the number of training IDs. * denotes unpublished papers.

Methods	rank-1	mAP
BoW+kissme [51]	25.13	12.17
LOMO+XQDA [24]	30.75	17.04
GAN (R, 702) [54]	67.68	47.13
B1 (R, 702)	64.22	43.50
B2 (R, 702)	52.91	31.23
APR (R, 702)	**70.69**	**51.88**

Table 2. Comparison with the state of the art on DukeMTMC-reID. Rank-1 accuracy (%) and mAP (%) are shown. Notations are the same with Table 1.

First, while it is expected B1 achieves good performance [52], we observe that B2 also yields decent accuracy, *e.g.*, a rank-1 accuracy of 49.76% using ResNet-50 on Market-1501. In fact, B2 only utilizes the attribute labels without the ID loss. This illustrates that attributes are capable of discriminating between different persons.

Second, by integrating the advantages in B1 and B2, our method exceeds both baselines by a large margin. For example, when using ResNet-50 and 651 training IDs, the rank-1 improvement over B1 and B2 is 12.47% and 33.22%, respectively on Market-1501. Consistent findings also hold for DukeMTMC-reID, *i.e.*, we observe improvement of 6.47% and 17.78% over B1 and B2 in rank-1 accuracy, respectively. This demonstrates the complementary nature of the two baselines, *i.e.*, identity and attribute learning. In addition, a minor finding is that using more training IDs marginally increases the matching accuracy.

Figure 8. A sample re-ID result on the Market-1501 dataset. Images in red bounding boxes denote false matches.

Third, for both CaffeNet and ResNet-50, APR yields consistent improvement. On Market-1501 with 651 training IDs, the improvement in rank-1 accuracy is 5.41% and 12.47% on CaffeNet and ResNet-50, respectively.

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 Table 1 and Table 2, respectively. On Market-1501, we obtain **rank-1 = 84.29%, mAP = 64.67%** using the ResNet-50 model and 751 training IDs. We achieve the best rank-1 accuracy among the competing methods, and the second best in mAP (the highest mAP is reported by Gent *et al*. [9]). On DukeMTMC-reID, our results are **rank-1 = 70.69% and mAP = 51.88%** using ResNet-50 and the full training set (702 IDs). Our method is thus shown to compare favorably with the state-of-the-art methods. A group of sample re-ID results on the Market-1501 dataset is shown in Fig. 8. Baseline 1 fails to return any true matches in the top-8 images of the rank list. In B1, people with a *backpack* or of a different *gender* are retrieved. When using APR, all the six true matches are found. In this example, *bag* and *female* are the key attributes.

Results between camera pairs. To further understand the performance on the Market-1501 dataset, we provide the re-ID results between all camera pairs in Fig. 10. Although camera 6 is a 720 × 576 SD camera and captures distinct background with the other HD cameras, the re-ID accuracy between Cam-6 and the others is relatively high. The cross-camera average mAP and average rank-1 accuracy are 52.24% and 58.56%, respectively. Compared with the results reported in [51], our accuracy is significantly higher, and we also observe a smaller standard deviation between cameras, indicating that APR can work under various viewpoints.

Scalability of the learned representation. To test the scalability of our method, we report results on the Market-1501+500k dataset. The 500k distractor dataset is composed of background detections and a large number of irrelevant pedestrians. The re-ID accuracy of our model (ResNet, 751 training IDs) on this dataset is presented in

Fig. 11. It can be expected that the re-ID accuracy drops as the database gets larger due to the inclusion of more distractors. The results further show that our method outperforms both [53] and Baseline 1. Nevertheless, we notice that the gap between APR and B1 gets smaller as the gallery scales up, which is probably due to the transfer effect: the data distribution of the gallery deviates more from the training set when the 500k images are gradually added. So it remains challenging how to adapt the learned model in unseen testing galleries.

Ablation studies. We evaluate the contribution of individual attributes on the re-ID accuracy. We remove one attribute from the system at a time with a fixed $\lambda = 8$, and the results on the two datasets are summarized in Fig. 9. We find that for the 10 attributes on Market-1501 and the 9 attributes on DukeMTMC-reID, most of them 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 2.34% and 4.85% on the two datasets, respectively. This indicates that pedestrians of the two datasets have different appearances. The attribute of "wearing hat or not" seems to exert negative impact on the overall re-ID accuracy, but the impact is very small.

5.4. Evaluation of Attribute Recognition

We test attribute recognition on the galleries of the Market-1501 and DukeMTMC-reID datasets in Table 3 and Table 4, respectively. We compare the model learned by APR and Baseline 2. Two observations are made.

First, on both the Market-1501 and DukeMTMC-reID datasets, the overall attribute recognition accuracy is improved by the proposed APR network to some extent. The improvement is 0.69% and 0.06% on Market-1501 and DukeMTMC-reID, respectively. So overall speaking, the integration of identity classification introduces some degree of complementary information and helps in learning a more discriminative attribute model.

Second, we observe that the recognition rate of some attributes decreases for APR, such as *gender* and *boots* in DukeMTMC-reID. However, Fig. 9 demonstrates that these attributes are necessary in 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.

We show two examples of attribute prediction in Fig. 12. Our system makes correct predictions of all the attributes for the person on the left. For the person on the right, incorrect recognition is observed on the *long hair* and *wearing a hat or not* attributes.
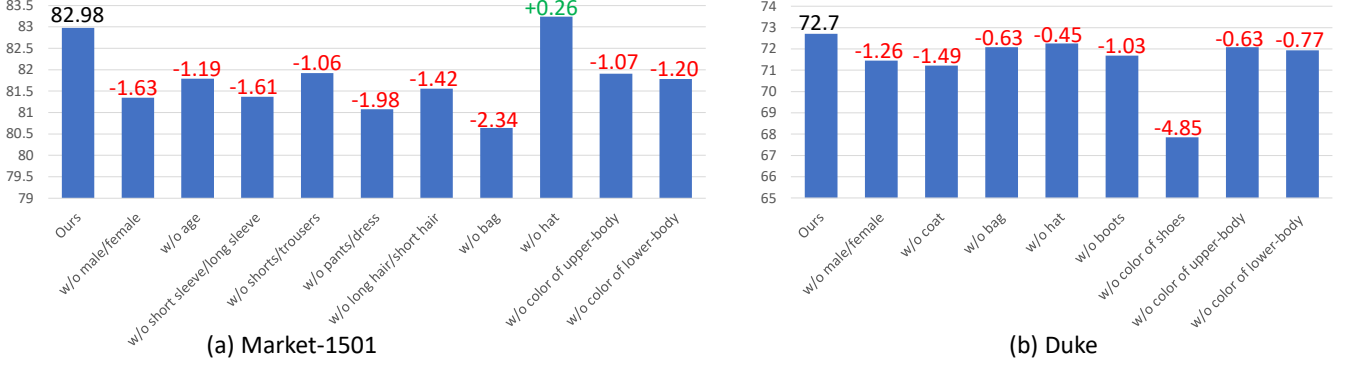
(a) Market-1501   (b) Duke

Figure 9. Re-ID rank-1 accuracy on Market-1501. 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.

Methods	gender	age	hair	L.slv	L.low	S.clth	B.pack	H.bag	bag	hat	C.up	C.low	mean
Baseline 2	86.63	84.35	81.85	93.50	91.69	93.98	84.63	86.74	76.23	97.06	72.66	66.37	84.64
APR	86.45	87.08	83.65	93.66	93.32	91.46	82.79	88.98	75.07	97.13	73.40	69.91	85.33

Table 3. 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*.

Methods	gender	hat	boots	L.up	B.pack	H.bag	bag	C.shoes	C.up	C.low	mean
Baseline 2	83.09	86.37	87.42	89.42	78.65	93.34	82.20	86.99	73.17	40.06	80.07
APR	82.61	86.94	86.15	88.04	77.28	93.75	82.51	90.19	72.29	41.48	80.12

Table 4. Attribute recognition accuracy on DukeMTMC-reID. "C.shoes" denote *color of shoes*, and the other notations are the same with Table 3. Note that λ is optimized on Market-1501.
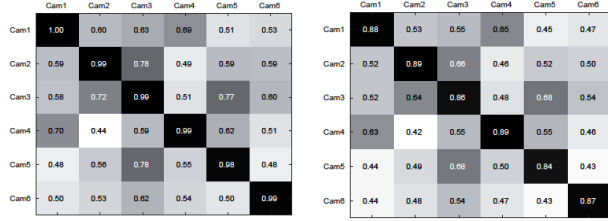

Figure 10. Re-ID performance between camera pairs on Market1501. (1) mAP and (2) rank-1 accuracy. Cameras on the vertical and horizontal axis correspond to the probe and gallery, *resp*. The cross-camera average mAP and average rank-1 accuracy are 52.24% and 58.56%, *resp*.
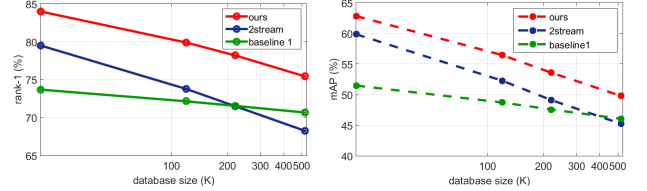

Figure 11. Re-ID accuracy on the Market-1501+500k dataset. (Left:) rank-1 accuracy. (Right:) mean average precision. We compare our method with [53] and Baseline 1.

6. Conclusions

In this paper, we mainly discuss how re-ID gets improved by the integration of attribute learning. The two tasks are to some extent mutually benefited from the multi-task learning process. We propose the attribute-person recognition (APR) network which learns a discriminative embedding for person re-ID and is able to make attribute predictions. The APR network contains both the ID clas-sification and attribute classification losses which are respectively contained in the re-ID and attribute recognition baselines. To demonstrate the effectiveness of our method, we have annotated attribute labels on two large-scale re-ID datasets. We show that the APR network brings improvement over the two baselines in re-ID accuracy. We report very competitive re-ID accuracy to the state-of-the-art approaches. For attribute recognition, the improvement is kind of mixed but still we observe overall precision improvement.

In the future, more investigations will be made into how attributes and re-ID help each other. Various attributes such

Female	3.04
Long hair	1.51
Short sleeve	2.15
Short bottom	2.70
Dress	1.05
No hat	2.43
Backpack	0.01
No handbag	0.04
Yellow top	-0.53
Black bottom	0.96

Female	3.10
Long hair	3.21
Short sleeve	2.41
Short bottom	3.37
Pants	0.12
No hat	3.51
No Backpack	0.54
Handbag	0.10
White top	1.78
Black bottom	0.99

Figure 12. Examples for person attribute recognition. The two tables show the predicted attributes and the classification scores. Red bounding boxes indicate incorrect predictions.

as localized or relative attributes [8, 30] will be studied.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2

[2] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *arXiv:1701.03153*, 2017. 2

[3] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011.

[4] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 6

[5] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[7] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, 2014. 3

[8] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012. 9

[9] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv:1611.05244*, 2016. 2, 6, 7

[10] G. Gkioxari, R. Girshick, and J. Malik. Actions and attributes from wholes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2470–2478, 2015.

[11] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.

[12] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[14] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[15] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. *arXiv:1603.00370*, 2016. 6

[16] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis. Joint learning for attribute-consistent person re-identification. In *ECCV*, 2014. 1, 3

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4

[18] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004. 3

[19] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014. 3

[20] R. Layne, T. M. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014. 3

[21] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *BMVC*, 2012. 1, 3

[22] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv:1603.07054*, 2016. 3

[23] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2

[24] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 6

[25] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[26] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV*, pages 391–401. Springer, 2012.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 3

[28] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *ICIP*, 2013. 3

[29] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *TPAMI*, 24(5):707–711, 2002. 3

[30] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 9

[31] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 5

[32] L. Shuang, X. Tong, L. Hongsheng, Z. Bolei, Y. Dayu, and W. Xiaogang. Person search with natural language description. *arXiv:1702.05729*, 2017. 3

[33] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *ICCV*, 2015. 1, 3

[34] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. *arXiv:1605.03259*, 2016. 1, 6

[35] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014. 2

[36] E. S. Tetsu Matsukawa. Person re-identification using cnn features learned from combination of attributes. *ICPR*, 2016. 3

[37] E. Ustinova, Y. Ganin, and V. Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *arXiv:1512.05300*, 2015. 6

[38] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2, 6

[39] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 6

[40] L. Wu, C. Shen, and A. v. d. Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *arXiv:1606.01595*, 2016. 6

[41] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv:1604.07528*, 2016. 2

[42] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv:1604.01850*, 2016. 2

[43] S. Yang, P. Luo, C.-C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *ICCV*, 2015. 3

[44] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014. 2

[45] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. *arXiv:1603.02139*, 2016. 6

[46] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.

[47] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014.

[48] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 3

[49] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 2

[50] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv:1701.07732*, 2017. 6

[51] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 3, 5, 6, 7

[52] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv:1610.02984*, 2016. 1, 2, 4, 6

[53] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *arXiv:1611.05666*, 2016. 2, 6, 7, 8

[54] Z. Zheng, L. Zheng, Y. Zheng, and Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv:1701.07717v3*, 2017. 1, 2, 3, 5, 6