A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios

Dangwei Li, *Student Member, IEEE*, Zhang Zhang, *Member, IEEE*, Xiaotang Chen, *Member, IEEE*, and Kaiqi Huang, *Senior Member, IEEE*

***Abstract*— Retrieving specific persons with various types of queries, e.g., a set of attributes or a portrait photo has great application potential in large-scale intelligent surveillance systems. In this paper, we propose a richly annotated pedestrian (RAP) dataset which serves as a unified benchmark for both attribute-based and image-based person retrieval in real surveillance scenarios. Typically, previous datasets have three improvable aspects, including limited data scale and annotation types, heterogeneous data source, and controlled scenarios. Differently, RAP is a large-scale dataset which contains 84 928 images with 72 types of attributes and additional tags of viewpoint, occlusion, body parts, and 2589 person identities. It is collected in the real uncontrolled scene and has complex visual variations in pedestrian samples due to the change of viewpoints, pedestrian postures, and cloth appearance. Towards a high-quality person retrieval benchmark, an amount of state-of-the-art algorithms on pedestrian attribute recognition and person re-identification (ReID), are performed for quantitative analysis with three evaluation tasks, i.e., attribute recognition, attribute-based and image-based person retrieval, where a new instance-based metric is proposed to measure the dependency of the prediction of multiple attributes. Finally, some interesting problems, e.g., the joint feature learning of attribute recognition and ReID, and the problem of cross-day person ReID, are explored to show the challenges and future directions in person retrieval.**

Manuscript received June 4, 2018; revised September 6, 2018 and October 14, 2018; accepted October 16, 2018. Date of publication October 26, 2018; date of current version November 28, 2018. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001005, in part by the National Natural Science Foundation of China under Grant 61473290 and Grant 61673375, and in part by the Projects of Chinese Academy of Science under Grant QYZDB-SSW-JSC006 and Grant 173211KYSB20160008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tolga Tasdizen. *(Corresponding author: Kaiqi Huang.)*

D. Li and X. Chen are with the Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: dangwei.li@nlpr.ia.ac.cn; xtchen@nlpr.ia.ac.cn).

Z. Zhang is with the Center for Research on Intelligent Perception and Computing and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zzhang@nlpr.ia.ac.cn).

K. Huang is with the Center for Research on Intelligent System and Engineering and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China, and the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: kqhuang@nlpr.ia.ac.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a detailed description about the RA dataset, the attribute recognition results at different environmental and context factors, and all the query items used in attribute-based person retrieval. The total size of the files is 0.05 MB. Contact kqhuang@nlpr.ia.ac.cn for further questions about this work.

Digital Object Identifier 10.1109/TIP.2018.2878349

***Index Terms*— Pedestrian retrieval, person re-identification, pedestrian attribute recognition, multi-label learning.**

I. Introduction

PERSON retrieval with the querying conditions of specific visual attributes or portrait images is very useful in hunting criminal or terrorism suspects in large-scale surveillance scenarios. For example, describable person attributes can play the critical role in the retrieval of the two suspects in Boston marathon bombing event [1]. To complete the task, how to extract a "good" feature representation for the target person is a crucial and challenging problem due to the low image quality, the large variations of camera viewpoints, the large pose variations and occlusions in real unconstrained scenes. Although deep neural network based methods learning visual features from large-scale training samples have achieved a series of breakthroughs in various vision tasks, the used large-scale benchmark datasets e.g. ImageNet and COCO, are collected from Internet, which have intrinsic bias of cyberspace, e.g. selection bias, capture bias, and negative set bias [2]. Thus, for person retrieval in the domain of physical world, it is necessary to construct a large-scale and richly annotated pedestrian dataset for feature learning and algorithms evaluations. In this paper, considering two types of query modalities in person retrieval, i.e., image-based query and attribute-based query, as shown in Fig. 1, we collect a large-scale and richly annotated pedestrian (RAP) dataset as a unified benchmark for person retrieval in real visual surveillance scenarios.

For person retrieval with the image-based query, ReID techniques which aim to find the target persons from a large-scale gallery database, have drawn a lot of attention in the past decade. It assumes that the target person has at least one image in the gallery who has the same identity and been captured by different cameras from that of probe image. Currently, feature extraction and similarity calculation are two main steps for ReID, where many powerful features of person images and similarity metrics have been proposed [3]. Since ReID aims to retrieve person instances whose identities are exactly the same as that in the query image, and the test persons are never seen during training, so that ReID can be stated as one kind of zero-shot learning [4]. In complex surveillance scenarios, the ReID model is often very hard to acquire satisfying results. Thus, an alternative relaxed objective to retrieve persons with certain attribute conditions may be easier achieved.

Attribute-based person retrieval searches interesting persons using describable pedestrian attributes, e.g., gender, and clothing types. Typically, for the attribute-based person retrieval,

TABLE I

COMPARISON AMONG DIFFERENT PEDESTRIAN ATTRIBUTE DATASETS. "PID" MEANS THE ANNOTATION IS BASED ON PERSON IDENTITY, WHILE THE "PI" MEANS THE ANNOTATION IS BASED ON EACH PERSON INSTANCE. "BATTRIBUTES" REPRESENTS THE BINARY ATTRIBUTES

Datasets	#Cams	Scene	Type	#Samples	Resolution (W×H)	#Battributes	Viewpoint	Occlusion	Part
VIPeR [10]	2	outdoor	PID	1264	48×128	21	yes	no	no
PRID [11]	2	outdoor	PID	400	64×128	21	no	no	no
GRID [12]	8	outdoor	PID	500	from 29×67 to 169×365	21	no	no	no
APiS [6]	-	outdoor	PI	3,661	48×128	11	no	no	no
Market-1501 [8]	6	outdoor	PID	32,668	64×128	13	no	no	no
DukeMTMC-reID [9]	8	outdoor	PID	34,213	from 34×85 to 193×477	8	no	no	no
PETA [5]	-	mixture	PID	19,000	from 17×39 to 169×365	61	no	no	no
RAP	25	indoor	PI	84,928	from 33×81 to 415×583	69	yes	yes	yes

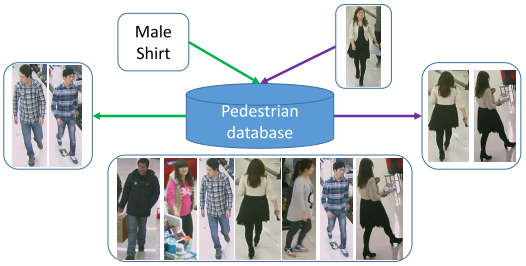


Fig. 1. The general framework for person retrieval based on different types of queries. The green and purple lines represent attribute-based and image-based queries, respectively.

Fig. 2. Image samples in RAP. In the real scene, attributes will change or be hard to determine due to camera viewpoint, body part occlusion, human's orientation and pose, time range, image quality etc. Even the same person's attributes will change a lot, which makes it a challenging problem.

TABLE II

COMPARISON AMONG DIFFERENT REID DATASETS. "#ID", "#BBOX", "#DIST" AND "#CAM" REPRESENT THE NUMBER OF IDENTITIES, BOUNDING BOXES, DISTRACTORS AND CAMERAS, RESPECTIVELY

Datasets	#ID	#Bbox	#Dist	#Cam	Label	Metric
VIPeR [10]	632	1,264	0	2	hand	CMC
i-LIDS [19]	119	476	0	2	hand	CMC
CUHK01 [20]	971	3,884	0	2	hand	CMC
CUHK02 [21]	1,816	7,264	0	2	hand	CMC
CUHK03 [7]	1,360	13,164	0	2	DPM & hand	CMC
DukeMTMC-reID [9]	1,404	34,213	0	8	Doppia [22]	mAP
MARS [13]	1,261	1,191,003	3,248	6	DPM & GMMCP	mAP& CMC
Market-1501 [8]	1,501	32,668	2,793 +500K	6	DPM & hand	mAP
RAP	2,589	26,638	14,947	23	hand	mAP& CMC

we usually train classifiers for all the predefined attributes based on training data firstly. In the test stage, the likelihoods of the gallery images to the query attribute categories are calculated. At last, the ranking list based on attribute likelihoods are returned. Although attribute-based person retrieval has a more relaxed objective than ReID, and has many technical advantages such as low storage cost and high retrieval efficiency, it is still hard to obtain satisfying results due to its inherent challenges, e.g., the large intra-class variations in attribute categories (appearance diversity and ambiguity [5]), extremely unbalance distribution and lack of sufficient training data.

As far as we know, existing datasets typically tend to one type of retrieval, either attribute or image. For the attribute-based person retrieval, there are two popular datasets, i.e., the APiS [6] and the PETA [5]. The APiS has fewer images and annotation types, which only has 3,661 images with 11 binary attributes and 3 multi-class attributes. Although the PETA has a larger number of images and annotation types, it is annotated at identity level, i.e., person images belonging to the same identity have the same annotations, even some attributes are invisible due to some factors, like occlusion or viewpoints. For the image-based person retrieval, the current large-scale datasets are CUHK03 [7], Market1501 [8] and DukeMTMC-reID [9]. Although they have a large number of identities, e.g. 1,501 identities of Market1501, they are still limited in the intra-class variations, such as occlusions, environments, and partial variations of cloth appearance.

In this paper, we collect a new large-scale Richly Annotated Pedestrian (RAP) dataset to accelerate the research of person retrieval in real surveillance scenarios. Some typical samples are shown in Fig. 2. The RAP has totally 84,928 pedestrian samples collected from 25 scenarios, and each sample is annotated with 72 fine-grained attributes as well as viewpoints, occlusions, and body parts. In addition, person identities are annotated in a subset of RAP, which contains 41,585 images, for the study of image-based person retrieval. It consists of 2,589 identities who appears at more than two cameras, and extra 14,947 person images as distractors. The comparisons between RAP and existing pedestrian attribute datasets and ReID datasets are shown in Table I and Table II. Compared with other attribute datasets, RAP has more data samples, fine-grained attributes, and extra contextual annotations, e.g. time, locations, viewpoints, occlusion patterns, and body parts. Compared with other ReID datasets, there are more identities, as well as *instance-level* attribute annotation, which may be useful in developing new attribute assisted ReID models. In summary, RAP[1] provides a unified evaluation benchmark

[1] http://rap.idealtest.org/, will be updated soon.

on large-scale person retrieval based on two query types, i.e., person attributes and identities, in unconstrained real scenarios.

The contributions of this paper are summarized as follows.

– A large-scale richly annotated pedestrian dataset RAP is collected to accelerate the research on person retrieval in real surveillance scenarios. To the best of our knowledge, RAP is current largest person retrieval dataset which could support pedestrian attribute recognition, attribute-based person retrieval, and person ReID simultaneously.
– Standard data splits and evaluation metric are used and several state-of-the-art algorithms are conducted on RAP, which make it available to evaluate and develop new approaches on pedestrian retrieval. Furthermore, instance-based metrics in multi-label learning are firstly introduced in attribute classification, which could be better to evaluate the dependency of multiple attributes on each sample.
– For pedestrian attribute classification, the influences of viewpoints, occlusion styles, and body parts are empirically analyzed, which can inspire the development of advanced algorithms in future.
– Several experiments are conducted to explore the relationship between attribute classification and ReID, which may help the model design in future. Moreover, for ReID based person retrieval, a more challenging problem of cross-day person ReID is explored, where the limitations of current ReID approaches are discussed.

The rest of this paper is organized as follows. Section II reviews related work. Section III describes the collected RAP database and evaluation protocols. Section IV introduces the baseline methods. Experiments are shown in Section V. In the last, Section VI concludes the paper.

II. Related Work

A. Related Datasets

In a large-scale surveillance system, person retrieval is particularly useful for security officers/operators searching for terrorists or suspects, which can be performed by attribute-based or image-based queries. As the core technique of image-based person retrieval, ReID has become one hot topic recently. Some large-scale datasets have been proposed, e.g. Market-1501 [8], DukeMTMC-reID [9], and MARS [13]. Market-1501 has 1,501 identities collected from 6 manually settled cameras in the campus. Based on the same data source of Market-1501, MARS is collected for the study of video-based ReID, which has 1,261 identities. Different from Market-1501 and MARS, DukeMTMC-reID is collected from uncontrolled campus scenes, and it has 1,404 identities from 8 cameras.

Visual attributes have been proved to be advantageous for image classification and retrieval [14], [15]. The first public pedestrian attribute dataset, i.e. the APiS dataset [6] is presented at the year 2013, which has 3,661 samples with 11 binary attributes. Deng *et al.* [5] collect a larger pedestrian attribute dataset PETA in 2014, which has 64 binary attributes and 19,000 samples. Recently, Li *et al.* [16] propose a language-based person search dataset, which aims to learn language-image matching for person retrieval.

Instead of studying ReID and pedestrian attribute independently, recently researchers also pay more attention on the influence of attributes on ReID [17], [18]. Popular ReID datasets, such as Market-1501, are also annotated with extra attributes. However, the additional attribute annotations has limited attribute categories, where only more than ten binary attributes are annotated. Furthermore, the attribute are annotated based on person identity, where all image samples owning the same ID are assigned with the identical attribute value, regardless that the visibility of the attributes may be changed, such as handbag and glasses, due to the variance of camera viewpoints, person postures, and occlusion conditions.

Different from previous datasets which are oriented to either attribute recognition, or ReID, RAP is constructed for a unified evaluation benchmark on large-scale person retrieval with both types of queries, i.e., attribute-based and image-based queries. The simultaneous annotation of instance-level attributes and ReID relationships also provides a richly annotated data platform for developing advanced algorithms on pedestrian recognition, i.e., attribute assisted person ReID.

B. Pedestrian Attribute Recognition

Previously, pedestrian attribute recognition is studied based on handcrafted features with traditional classifiers. Layne *et al.* [17], [23] use low-level features and Support Vector Machines (SVM) [24] to detect attributes. Deng *et al.* [5] utilize the kernel SVM [25] to recognize attributes, where the Markov Random Field is used to learn the relationship among attributes. Zhu *et al.* [6] introduce Gentle AdaBoost to accomplish feature section and classifier learning jointly.

Inspired by the great success of Deep Convolutional Neural Networks (DCNN) in image recognition [26], several DCNN-based methods are proposed in pedestrian attribute recognition. Typically these methods learn image features and classifier in an end-to-end style. Li *et al.* [27] propose a Multi-Attribute Recognition (DeepMAR) model, which utilizes the prior knowledge in the objective function for attribute recognition. Zhang *et al.* [28] propose a pose aligned networks for deep attribute modeling and use the poselets [29] to assist attribute classification in the natural scene. Zhu *et al.* [30] introduce the patch-based multi-label CNN with predefined attribute-patch connection structure to recognize attributes. Sudowe *et al.* [31] propose the Attribute Convolutional Net (ACN) to learn multiple attributes through a jointly-trained holistic CNN model. In this paper, we adopt two state-of-the-art models, e.g. CaffeNet [26] and ResNet [32], with the objective function of DeepMAR and ACN for pedestrian attribute recognition task.

C. Attribute-Based Person Retrieval

Attribute-based person retrieval is first proposed by Vaguero *et al.* [14], where the parsing of human parts and their attributes, e.g. facial hair, clothing color, are studied for person retrieval. After that, many approaches are proposed for person retrieval with attribute-based queries [1], [33]–[35]. Thornton *et al.* [33] propose to use a generative probabilistic

model to match the queried attributes and gallery images. Feris *et al.* [1] utilize the "learning to rank" approach to fuse the score of different attributes' detection scores. Shi *et al.* [34] transfer the semantic representation from the human parsing dataset to pedestrian images in surveillance for person retrieval. When querying with multiple attributes, the product of probabilities of multiple attributes are used to generate the ranking list. In this paper, we adopt a two-step strategy for attribute-based person retrieval, where pedestrian attributes are firstly recognized, then similarity ranking of multiple attributes is calculated with probability-product principle [34].

D. Image-Based Person Retrieval

Image-based person retrieval, where the core technique is ReID, has made great progress in past few years [3], [36]. Most approaches focus on developing powerful features to handle problems like the variations of viewpoint [37], body pose [38], illumination [39], appearance [40], or learning an effective distance metric [41]–[43]. With the increasing sample size of ReID database, features learning from multi-class person identification with CNN [13], [38], [44], denoted as ID-discriminative Embedding (IDE) [45], has shown great potentials on large-scale ReID datasets, such as MARS [13], PRW [45], where the IDE features are obtained from the last hidden layer of DCNN.

In addition to image-based person ReID, there are some work utilizing describable person attributes, e.g. gender, cloth types, to assist ReID, such as [17], [18], [46], and [47]. Although binary attributes have great potentials in ReID [17], they are hard to be directly used due to the weak detection results. Typically, researchers make use of attributes in two ways. First, the detection scores of all the attributes are extracted as extra attribute features. The similarities between attribute features are computed, which are further fused with other types of similarities for ReID, e.g. [17], [46]. Second, attributes are used to assist the learning of robust semantic features which are further used in ReID, e.g. [18], [47]. In this paper, we focus on the second one and explore the richly annotated information in RAP to jointly learn the IDE representation with the supervision of person attributes and identity.

III. The RAP Dataset

A. Data Collection and Annotation

1) Data Collection: RAP is collected from a realistic High-Definition (1,280×720) surveillance network at an indoor shopping mall. Considering the data richness and storage cost, totally 30 days from 25 cameras at 15 frames per second are collected firstly. Then a motion detection and tracking algorithm based on Gaussian Mixture Model is applied to all the collected videos on a ten nodes MPI cluster. For each camera, one or two virtual lines indicating region-of-interest are set manually and the tracked objects accessing the virtual lines are used for annotation. Considering the annotation cost, totally 30 hours (1 hour per day) synchronous videos are selected for each visual scene for annotation. Finally, 84,928 human images are used to construct the RAP dataset.

TABLE III
DETAILED ANNOTATIONS IN RAP. COMPLETE ATTRIBUTE NAMES ARE SHOWN IN SUPPLEMENTAL MATERIALS

Class		Attributes
Spatial-Temporal		time, sceneID, bounding box of body/head-shoulder/upper-body/lower-body/accessories.
Whole		gender, age, body shape, role.
Accessories		backpack, single shoulder bag, handbag, plastic bag, paper bag etc.
Postures, Actions		viewpoints, telephoning, gathering, talking, pushing, carrying etc.
Occlusions		occluded parts, occlusion types.
Parts	Head	hair style, hair color, hat, glasses.
	Upper	clothes style, clothes color.
	Lower	clothes style and color, footware style and color.

2) Attribute: To cover most important attributes in a practical surveillance system, 72 attributes are adopted, including some attributes that are suggested by the UK Home Office and UK police to be the most valuable in tracking and criminal identification [48]. Besides, the viewpoints, occlusions, and body part positions are annotated to study the effect of the contextual and environmental factors for person retrieval. The annotation is carried out by a professional data corporation and the quality of annotation is guaranteed by 3 rounds of sample checking as well as the workers and our staffs. Detailed descriptions of the annotations are shown in Table III.

As shown in Table I, in RAP, each human image is annotated independently and the person images of the same identity may have different attribute annotations due to the viewing angle variations, occlusions, and the large time interval of occurrences etc. This is especially important in indoor surveillance scenes where a huge number of pedestrian samples have some occlusions in various degrees, e.g., about 28% images are occluded in RAP. In addition, RAP is collected during a continuous period with 25 cameras scenes in a unified surveillance network, which provides sufficient variations in person images, e.g., postures, view angles, etc., but less scenario heterogeneity than that of PETA dataset which is derived from the mixture of different surveillance scenes with different image qualities and data environments.

3) Person ID: Besides attributes, we manually select 41,585 images over 15 days from entire RAP for identity annotation. Finally, there are 2,589 identities with 26,638 samples are annotated for ReID. The rest 14,947 pedestrian samples serve as distractors who have disjoint identities. Compared with other ReID datasets, RAP has more identities and cameras. The distractors are pedestrian images rather than pure backgrounds, which could increase the difficulty of person retrieval. In addition, the identities are also annotated with extra information, e.g. viewpoints, occlusion patterns, body parts, which will help to develop better ReID algorithms.

In summary, RAP contains more samples, annotation categories and variations in camera views, and provides a unified benchmark to demonstrate the effectiveness of the various approaches on person retrieval with both types of queries.

TABLE IV

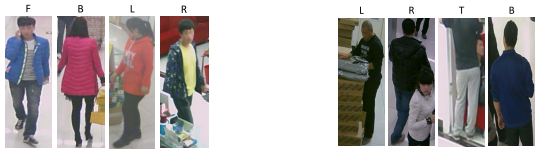
DISTRIBUTION OF PEDESTRIAN VIEWPOINT. THE PERCENTAGES ARE THE RATIO OF THE NUMBER OF IMAGES IN EACH VIEWPOINT TO THE WHOLE DATASET

Front (F)	Back (B)	Left (L)	Right (R)
19,678 (23%)	20,651 (24%)	22,039 (26%)	22,560 (27%)

TABLE V

DISTRIBUTION OF OCCLUSION POSITIONS. A PEDESTRIAN IMAGE MAY BE OCCLUDED BY MULTIPLE POSITIONS

Top (T)	Bottom (B)	Left (L)	Right (R)	Total
4,480	16,440	3,370	2,483	24,115



(a) (b)

Fig. 3. (a) and (b) show some examples of different viewpoints and occlusion patterns respectively. The viewpoints from left to right in (a) are front, back, left and right. The occlusion sources of first two images in (b) are attachment and person. The last two images in (b) are occluded due to camera borders.

(a) (b)

Fig. 4. (a) shows examples of three body parts. The green bounding box in the first column is the attachment. (b) shows examples of the fine-grained attributes. Best viewed in color.

TABLE VI

DISTRIBUTION OF THE NUMBER OF ATTRIBUTES ACCORDING TO THE PROPORTION OF POSITIVE SAMPLES TO THE WHOLE DATASET

Positive proportion	(0, 0.1)	[0.1, 0.2)	[0.2, 0.3)	[0.3, 1)
Number of attributes	49	5	8	7

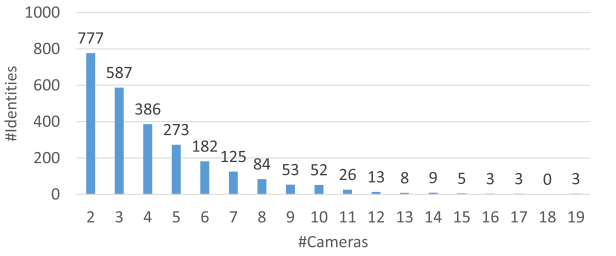


Fig. 5. Distributions of the number of person identities on the number of cameras that a person appears.

B. Dataset Characteristics

1) Attribute: Typically, RAP provides four extra kinds of annotations, namely viewpoints, occlusion patterns, body parts and fine-grained attributes.

a) Viewpoints: There are four types of viewpoints annotated in RAP, including facing front (F), facing back (B), facing left (L) and facing right (R). The viewpoint is annotated based on the full body's direction relative to the annotators. The distribution of different viewpoints is shown in Table IV. Examples of different viewpoints are shown in Fig. 3(a).

b) Occlusions: Occlusions happen very often in realistic scenes, especially in an indoor shopping mall. About 28% pedestrians in RAP have occlusions in various degrees. We further label the occluded samples according to the locations of occlusions and the sources of occlusions. The locations of occlusions are divided into 4 categories, i.e. top (T), bottom (B), left (L) and right (R). The occlusion is annotated when its corresponding parts are invisible. The distribution of occlusion locations is shown in Table V. There are four sources of occlusions, including persons, environments (such as camera border), attachments (such as package) and other. Different from the viewpoint, occlusion pattern can be multi-value due to the complex environment or contextual factors. Examples of occlusion patterns are shown in Fig. 3(b).

c) Body parts: Typically, humans in surveillance videos are relatively small and blur. Exact part positions, such as the arm and foot, are hard to be annotated in practical systems. However, body parts are essentially useful for variant tasks, such as pose estimation and ReID. In this paper, instead of fine-grained body part annotations, three coarse-grained body parts are annotated, including head-shoulder, upper body, and lower body, which can be used to develop more efficient part-based attribute or identity representation as well as study some new problems, such as attribute localization in future. Examples of the three parts are shown in Fig. 4(a).

d) Fine-grained attributes: Besides the environmental and contextual factors, we also annotate various interesting fine-grained attributes, including action (such as Talking and Pushing), role (such as Customer and Clerk), and body type (such as Fatter and Thinner). Example images are shown in Fig. 4(b).

To have a deeper analysis about attribute distributions, we provide the statistics of annotations. The number of attributes on one person ranges from 4 to 26, and the mean value is 12.2. Moreover, we also find that RAP has serious unbalance distributions measured by the proportion of positive samples to the whole dataset on most binary attribute categories. As shown in Table VI, most attributes have fewer positive samples (less than 10%), which makes it a challenging problem to develop high-quality recognition algorithms.

2) Identity: RAP has 2,589 person identities with 26,638 samples. Since the camera network is large, it is hard for a person walking across all cameras. The distribution of person identities *vs.* the number of cameras is shown in Fig. 5. In summary, a person would appear at 2 to 19 cameras and the mean value is 4.2. As shown in Table VII, another challenge is that the number of samples per identity has large variances and most identities have less than 5 samples. It is reasonable because most customers have specific purposes in the shopping mall, and they will occur in certain regions.

TABLE VII
DISTRIBUTION OF THE NUMBER OF IDENTITIES ACCORDING TO THE NUMBER OF IMAGES FOR EACH IDENTITY

Samples per identity	≤ 5	(5, 10]	(10, 20]	(20, 30]	>30
Number of identities	1,380	622	118	104	132

C. Evaluation Protocols

a) Pedestrian attribute recognition: The first protocol to evaluate the attribute recognition algorithms is the *mean accuracy* (*mA*), which could handle the unbalance attribute distribution. It has been adopted to evaluate the pedestrian attribute recognition [5] and face attribute recognition [49]. Considering the unbalance distribution of attributes, for each attribute, *mA* computes the classification accuracy of positive samples and negative samples separately and then take the average of them as the result of the attribute. After that, *mA* takes an average over all the attributes as the final result. The evaluation criterion can be formally calculated as follows.

$$mA = \frac{1}{2L}\sum_{i=1}^{L}(TP_i/P_i + TN_i/N_i) \tag{1}$$

where L is the number of attributes; P_i and TP_i are the numbers of positive samples and correctly predicted positive samples, respectively; N_i and TN_i are the number of negative samples and correctly predicted negative samples, respectively.

The above *label-based* solution, treats each attribute independent and ignores the inter-attribute dependency, which exists naturally in multi-attribute recognition. Considering this problem, besides *mA*, we propose to use the *instance-based* metrics to evaluate attribute recognition algorithms. *Instance-based* evaluation captures better the consistency of the predictions of multiple attributes in one image [50], and it includes four metrics: accuracy, precision, recall rate, and F value, which are defined as follows.

$$\begin{aligned} Accuracy_{inst}(f) &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \bigcap f(x_i)|}{|Y_i \bigcup f(x_i)|} \\ Precision_{inst}(f) &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \bigcap f(x_i)|}{|f(x_i)|} \\ Recall_{inst}(f) &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \bigcap f(x_i)|}{|Y_i|} \\ F_{inst}(f) &= \frac{Precision_{inst} * Recall_{inst}}{Precall_{inst} + Recall_{inst}} \end{aligned} \tag{2}$$

where N is the number of samples. For the i-th sample, Y_i is the ground truth positive labels, $f(x_i)$ returns the predicted positive labels. $|\cdot|$ is the set cardinality. F_{inst} is an integrated version of $Precision_{inst}$ and $Recall_{inst}$.

b) Attribute-based person retrieval: As an image retrieval problem, we can use mean average precision (mAP) to evaluate the attribute-based retrieval. For each given query, e.g. multiple attributes, the average precision is defined as follows.

$$AP = \frac{\sum_{r=1}^{N} P(r) \times rel(r)}{K} \tag{3}$$

where $P(r)$ is the precision at Rank-r, $rel(r)$ is a binary function that judges whether the image at Rank-r is revelent to the give query attributes, e.g. owing all the attributes in the query attribute list, and K is the number of related person images. AP is also known as the area under the precision-recall curve. Finally, we can obtain the mAP by averaging all the queries' average precisions.

c) Person re-identification: We use the popular metrics, including mAP and Rank-1 accuracy [8] to evaluate the ReID performance. The computation process of AP is the same as Eq. (3) except for the images from the same identity and camera as the query image. After all the queries' APs are obtained, we obtain the mAP through computing the mean value. The Rank-1 accuracy in Cumulated Matching Characteristics (CMC) curve is a complementary result.

IV. METHODS

A. Pedestrian Attribute Recognition

We adopt SVM and end-to-end CNN approaches as pedestrian attribute recognition baselines.

1) SVM: SVM has been used for pedestrian attribute recognition before [5], [17]. We adopt the open-source liblinear package[2] in this paper. The cost parameter C is selected from $\{0.01, 0.1, 1, 10, 100\}$ based on the validation set. Due to the serious unbalance distribution, not all the examples are used for training. Similar with Layne *et al.* [17], we randomly downsample the negative samples to the size of positive samples if positive samples are less than negative samples. The same rule is applied to the positive samples. Subsequently, the sampled subset is used to train the SVM. We adopt two types of features for attribute recognition, including popular Ensemble of Localized Features (ELF) [37] and off-the-shelf CNN features. The open source code[3] is used to extract ELF. For off-the-shelf CNN features, the "FC6" and "FC7" in CaffeNet and "Pool5" in ResNet50, which are both pretrained in ILSVRC-2012 1000-category classification task, are adopted. Notably, both of these extracted features are L2 normalized in the training and test stage.

2) End-to-End CNN: Besides the two-stage methods, the end-to-end deep learning methods have made great progress in attribute recognition [27], [28], [31], which jointly learn features and classifiers. In this paper, we adopt two attribute recognition models as end-to-end CNN baseline methods, including ACN [31] and DeepMAR [27]. ACN adds an extra fully connected layer for each attribute and uses Kullback-Leibler divergence loss to optimize the network. It has achieved good performance in Parse-27k [31]. DeepMAR proposes to use weighted sigmoid cross-entropy loss to optimize the network and has achieved state-of-the-art results in PETA. For DeepMAR, besides CaffeNet, we also explore the deeper ResNet50 as the steam network for attribute recognition.

The stochastic gradient descent (SGD) is adopted to optimize the networks. For CaffeNet-based models, we use the initial learning rate 1×10^{-3} and weight decay with 5×10^{-4}.

[2] http://www.csie.ntu.edu.tw/ cjlin/liblinear/
[3] https://github.com/rlayne/ELF-v2.0-Descriptor/

The learning rate will be divided by 10 after each 20K iterations with batch size of 256. The mean value subtraction, random mirror and crop (256×256 to 227×227) are adopted. The model at 50K iterations is adopted for evaluation, and the center crop is used at test stage. For DeepMAR with ResNet50, the learning rate will be divided by 10 after each 30K iterations with batch size of 32. For image preprocessing, the mean value subtraction, random mirror and image resize (224×224) are adopted. The model at 90K iterations is used for evaluation. Note that the CaffeNet and ResNet50 are pre-trained on the 1.2M images from the ILSVRC-2012 1000-category classification task and then finetuned on RAP.

B. Attribute-Based Person Retrieval

Due to the imperfect pedestrian attribute classification algorithms, researchers usually use the predicted scores instead of binary labels for searching [14], [34], [51]. As different algorithms can lead to different person retrieval results, in this paper, we adopt the attribute classification algorithms in Section IV-A, including SVM, ACN and DeepMAR. For the multi-attribute query, we use the probability-product of all the query attributes as the similarity for final person retrieval. The similarity between query attribute set Q and gallery image x can be formalized as follows:

$$S(Q, x|\Theta) = \prod_{j=1}^{M} prob(Q_j|x, \Theta) \tag{4}$$

where Θ is the model parameters, x is the test pedestrian sample, and Q is a set of query key-words, which contains M attributes. $prob(Q_j|x, \Theta)$ represents the prediction probability of x belonging to attribute Q_j given the model parameter Θ. Here we consider each attribute as an independent variable, thus the joint probability can be computed through the probability-product principle.

C. Image-Based Person Retrieval

Two types of baselines, including hand-crafted features with metric learning algorithms and deep feature learning approaches, are adopted for ReID. For the first category, four types of features and three metrics are adopted. The features include ELF [37], Local Maximal Occurrence (LOMO) [42], Gaussian Of Gaussian (GOG) [39] and JSTL-DGD [44]. The metrics include Euclidean metric, Keep It Simple and Straightforward (KISS) [41] and cross-view Quadratic Discriminant Analysis (XQDA) [42]. By combining different features and metric learning approaches, we naturally have 12 different baselines.

For the second category, several state-of-the-art methods, including CaffeNet [26], ResNet [32], DenseNet [52], PAR [18], MSCAN [38], MuDeep [53], HACNN [54], are implemented. To explore the influence of attributes on ReID feature learning, we develop a multi-task network based on two backbone networks, including CaffeNet and ResNet50, as shown in Fig. 6, to jointly learn the multi-class person identification and pedestrian attribute classification. As we adopt two steam networks, the learned features for ReID

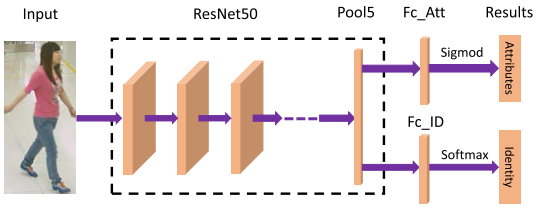


Fig. 6. The multi-task convolutional neural networks based on ResNet50 which jointly learn the attribute classification and multi-class identification tasks.

are 4,096-dimension "fc7" and 2,048-dimension "pool5" for CaffeNet and ResNet50, respectively. For multi-class person identification, the cross entropy loss with softmax is adopted as follows.

$$L_{id} = -\sum_{i=1}^{N} log \frac{\exp(W_{y_i}^T z_i + b_{y_i})}{\sum_{j=1}^{C} \exp(W_j^T z_i + b_j)} \tag{5}$$

where i is the index of person images, z_i is the feature of i-th sample, y_i is the identity of i-th sample, N is the number of images, C is the number of identities, W_j is the classifier for j-th identity. For attribute classification, we use the objective function of DeepMAR, which is defined as follows.

$$L_{att} = -\frac{1}{NL} \sum_{i=1}^{N} \sum_{l=1}^{L} w_l (y_{il} log(\hat{p_{il}}) + (1 - y_{il}) log(1 - \hat{p_{il}})) \tag{6}$$

where $w_l = \exp((1 - p_l)/\sigma^2)$ when $y_{il} = 1$, and $w_l = \exp(p_l/\sigma^2)$ when $y_{il} = 0$. y_{il} is the groundtruth of l-th attribute of i-th example, p_{il} is the corresponding prediction probability, w_l is the loss weight for l-th attribute, p_l is the positive ratio of l-th attribute in the training set, and σ is a temperature coefficient which is set as 1. The final objective function of the multi-task network is defined as follows.

$$L = L_{id} + \lambda L_{att} \tag{7}$$

where λ is the weight to balance two different losses. It is empirically set as 1 based on validation set.

The SGD is used to optimize the ReID backbone networks. The initial learning rate is 1×10^{-3} and decreases by 10 after every 20K iterations. The model at 50K iterations is used for evaluation. To avoid overfitting, Dropout [55] with rate 0.7 is adopted after the "fc6" and "fc7" layers in CaffeNet and "pool5" layer in ResNet50. For image preprocessing, the mean value subtraction, random mirror and random crop (256×256 to 227×227) are adopted in CaffeNet. In ResNet50, we resize the original image to 224×224, subtract the mean value in each channel, and random mirror the image for network training.

V. Experimental Results

This section will introduce the experiments of the above three tasks, i.e. pedestrian attribute recognition, attribute-based person retrieval, and image-based person retrieval.

TABLE VIII

ATTRIBUTE RECOGNITION RESULTS IN MA (%). "C" AND "R" REPRESENTS CAFFENET AND RESNET50 RESPECTIVELY. IACN AND IDEEPMAR ARE THE INDIVIDUAL ATTRIBUTE VERSIONS OF ACN AND DEEPMAR. DEEPMAR* ARE RESULTS BASED ON BODY PARTS. BEST RESULTS IN EACH ROW EXCEPT THE LAST TWO COLUMNS ARE INDICATED IN **BOLD**

Attributes	SVM-ELF	SVM-Fc6	SVM-Fc7	SVM-Pool5	IACN-C	ACN-C	IDeepMAR-C	DeepMAR-C	DeepMAR-R	DeepMAR*-C	DeepMAR*-R
Female	74.55	83.87	82.40	86.35	93.37	94.06	94.00	94.47	**96.53**	96.31	96.72
AgeLess16	71.51	82.87	81.41	**83.81**	70.80	77.29	77.49	79.45	77.24	81.71	80.36
Age17-30	62.23	63.92	63.66	65.57	67.17	69.18	67.88	68.08	**69.66**	72.07	70.60
Age31-45	60.37	61.36	61.20	62.89	64.60	**66.80**	65.01	65.12	66.64	68.88	67.41
Age46-60	64.38	68.50	68.53	**70.29**	50.23	52.16	58.48	57.41	59.90	66.68	58.63
BodyFat	59.24	63.91	63.41	**65.48**	55.41	58.42	63.60	61.92	61.95	67.24	61.54
BodyNormal	53.93	54.36	54.47	55.59	52.29	55.36	55.85	57.93	**58.47**	59.20	58.43
BodyThin	57.39	61.77	61.56	**62.46**	50.13	52.31	56.04	55.51	55.75	59.20	55.63
Customer	68.95	76.53	76.73	78.10	77.95	80.85	80.60	**82.68**	82.30	83.40	82.34
Employee	71.29	81.78	81.19	82.68	82.35	85.60	85.07	**86.74**	85.73	88.27	86.15
hs-BaldHead	68.94	75.66	72.56	74.88	56.27	65.28	65.38	69.91	**80.93**	86.83	83.46
hs-LongHair	71.60	82.36	80.84	84.93	89.20	89.49	90.64	90.28	**92.47**	92.56	91.89
hs-BlackHair	64.24	68.61	68.30	72.87	63.07	66.19	70.07	72.58	**79.33**	77.69	73.95
hs-Hat	64.95	75.20	73.28	75.78	54.11	60.73	63.56	67.47	**84.00**	83.98	82.06
hs-Glasses	63.87	66.13	66.32	67.48	60.31	56.30	70.53	69.67	**84.19**	84.62	86.02
ub-Shirt	67.99	74.43	73.46	76.21	80.77	81.81	83.04	83.61	**85.86**	85.53	86.05
ub-Sweater	59.25	63.68	63.11	**66.35**	53.55	56.85	62.12	64.33	64.21	65.82	64.11
ub-Vest	66.92	79.77	78.38	80.94	81.62	83.65	85.09	87.88	**89.91**	89.32	89.92
ub-TShirt	60.51	65.41	64.94	67.26	69.09	71.61	73.40	74.58	**75.94**	76.58	75.27
ub-Cotton	68.58	76.91	76.18	78.76	73.38	74.67	**79.23**	77.77	79.02	82.03	78.67
ub-Jacket	67.71	71.97	71.52	72.47	77.05	78.29	79.08	79.13	**80.69**	81.43	80.22
ub-SuitUp	67.87	77.68	77.41	**79.39**	70.07	73.92	77.29	77.56	77.29	82.83	79.28
ub-Tight	62.24	67.04	66.93	**69.80**	57.52	61.71	64.97	67.22	68.89	68.88	70.13
ub-ShortSleeve	73.43	84.19	82.90	85.11	87.19	88.27	89.56	89.66	**90.09**	91.44	89.37
ub-Others	57.43	61.62	61.54	**64.33**	50.00	50.35	52.62	54.04	54.82	55.09	53.86
lb-LongTrousers	73.35	76.31	74.77	76.88	85.60	86.60	85.84	86.10	**86.64**	88.67	87.89
lb-Skirt	69.31	80.14	78.86	**82.67**	63.81	70.51	73.06	74.23	74.83	79.48	75.82
lb-ShortSkirt	69.92	81.06	79.98	**82.63**	64.32	73.16	73.50	74.64	72.86	80.34	74.12
lb-Dress	69.17	82.54	80.85	**83.73**	67.19	72.89	76.75	77.05	76.30	83.55	75.70
lb-Jeans	81.56	81.22	79.15	82.74	90.04	90.17	**91.11**	89.68	89.46	91.73	90.04
lb-TightTrousers	71.33	83.09	81.43	84.72	84.88	86.95	**87.86**	87.07	87.91	91.38	88.81
shoes-Leather	67.04	69.82	68.90	71.24	71.27	71.92	74.93	74.91	**80.50**	79.44	81.11
shoes-Sports	63.16	66.81	65.59	68.72	62.21	62.59	69.84	67.88	**71.58**	73.68	71.88
shoes-Boots	69.72	81.95	80.25	82.86	83.55	85.03	87.48	87.79	**91.37**	92.64	91.28
shoes-Cloth	69.79	76.82	76.37	**77.89**	63.80	68.74	71.22	72.85	72.31	73.84	70.60
shoes-Casual	61.61	61.73	61.07	64.25	54.19	54.57	61.93	61.57	**64.58**	65.50	64.01
shoes-Other	64.35	66.70	66.76	**67.93**	51.17	52.42	60.83	59.12	61.56	66.10	61.39
att-Backpack	66.74	76.17	76.20	80.07	66.33	68.87	73.88	75.49	**80.61**	75.87	78.86
att-ShoulderBag	63.24	72.32	71.71	75.02	70.61	69.30	77.22	77.36	**82.52**	77.87	82.37
att-HandBag	63.69	71.74	71.43	74.72	60.64	63.95	70.33	72.44	**76.45**	70.40	76.58
att-Box	64.24	71.38	70.25	**76.92**	65.37	66.72	71.84	72.10	76.18	71.47	76.13
att-PlasticBag	60.54	70.30	69.89	73.32	62.62	61.53	69.40	68.91	**75.20**	69.62	77.89
att-PaperBag	61.12	69.37	67.98	**70.99**	50.45	52.25	56.70	58.64	63.34	56.33	62.18
att-HandTrunk	70.24	81.89	80.40	**85.46**	74.54	79.01	80.01	82.35	84.57	81.43	85.48
att-Other	58.82	62.58	62.19	65.25	67.13	66.14	70.13	70.65	**76.14**	70.49	76.79
act-Calling	63.21	73.13	70.82	74.77	77.93	74.66	83.54	81.67	**86.97**	85.58	87.23
act-Talking	62.29	65.83	65.38	**67.97**	50.00	50.54	54.25	54.70	54.65	53.96	53.66
act-Gathering	64.05	66.06	66.47	**70.30**	51.11	52.69	59.20	57.30	58.81	58.43	58.46
act-Holding	62.36	72.30	70.68	**73.30**	53.79	56.43	62.33	61.70	64.22	61.54	63.06
act-Pushing	72.38	86.36	85.37	**88.07**	77.64	80.97	80.87	84.72	82.58	83.37	81.62
act-Pulling	66.31	76.44	76.86	**79.79**	62.40	69.00	70.22	74.30	78.35	74.52	79.54
act-CarryByArm	57.57	65.99	65.30	**66.52**	51.82	53.55	60.64	60.21	65.40	59.67	65.32
act-CarryByHand	62.78	70.97	70.12	73.50	74.20	74.58	78.55	79.31	**82.72**	80.85	82.90
act-Other	63.36	71.12	70.51	**72.27**	51.74	54.83	59.68	58.89	58.79	59.43	56.27
Average	65.60	72.62	71.81	74.52	66.63	68.92	72.29	72.94	**75.54**	76.01	75.54

A. Person Attribute Recognition

1) Overall Evaluation: Two types of methods are implemented for attribute recognition. The first category is classic methods, where SVM classifiers with variant features (ELF, FC6, FC7, and Pool5) are adopted. The second category is end-to-end deep learning methods. It consists of two popular multi-attribute recognition models (ACN and DeepMAR), and their individual attribute versions (IACN and IDeepMAR). The experiments are conducted with 5 random splits where data partition is shown in Table IX. We use the validation set to select the proper hyperparameters, e.g. penalty factor C in SVM, and use the **train+val** to obtain the final model for the final evaluation on the test set. Due to the highly unbalanced distribution of attributes, totally 54 binary attributes are selected in our experiments. An attribute is selected when the proportion of positive samples to the whole dataset is higher than 0.01. Baldhead and Child are also selected due to their salience in real scenario. The results are shown in Table VIII.

TABLE IX

DATA SPLIT FOR PEDESTRIAN ATTRIBUTE RECOGNITION

Data partition	Train	Validation	Test
#Images	50,957	16,986	16,985

As shown in Table VIII, the attributes, e.g. gender and backpack, have better recognition accuracy, which is consistent with Deng *et al.* [5]. Compared with ELF, the off-the-shelf

Fig. 7. Recognition results of partial attributes in mA (%) at different viewpoints. Avg. of clean is the average results of all the clean test images without occlusion. Best viewed in color.

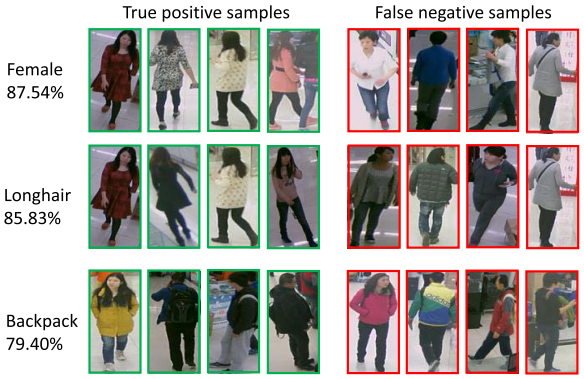


Fig. 8. Examples of partial attributes' recognition results. The green boxes are true positives and the red ones are false negatives. Best viewed in color.

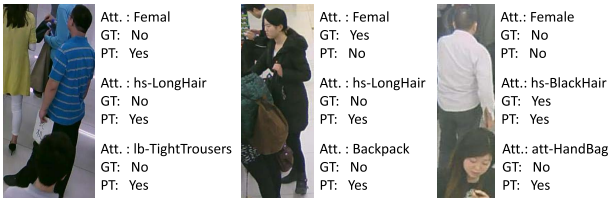


Fig. 9. Attribute recognition results of three samples from "person occluded by other person". Att. means pedestrian attribute category. GT. is the ground truth label. PT. is the predicted label. The first two samples show the false predictions and the third one shows the partial false predictions.

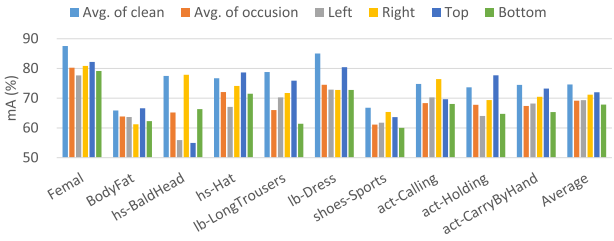


Fig. 10. Partial attributes' results in mA (%) on occlusion images. Avg. of clean is the average results of the clean test images without occlusion. Avg. of occlusion is the average results of the occlusion test images. Best viewed in color.

CNN features have shown better results. Compared with FC7, FC6 has better generalization ability. Compared with CaffeNet, the Pool5 from ResNet50 has shown better performance. Compared to the independent training strategy in IACN and IDeepMAR, the multi-attribute recognition models (ACN and DeepMAR) have shown better results on most of selected attributes. However, for a few attributes, e.g. hs-Galsses, att-PlasticBag, superior results can be obtained by the independent models, which suggests more appropriate multi-task learning strategies to learn shared features through efficiently exploring the relationships among multiple attributes. The *instance-based* evaluation will be shown on Table X for further comparison.

2) Influence of Viewpoint: To explore the influence of viewpoint, we manually select the images without occlusion for analysis. The clean training data with different viewpoints is used to train the linear SVM with Pool5 features. The test images are divided into four partitions according to the ground truth viewpoints, including front, back, left and right. Recognition results of selected partial attributes are shown in Fig. 7. In summary, partial attributes are sensitive to the viewpoint, such as Backpack. Besides the quantitative results, we also visualize the recognition results of three attributes, including Female, Longhair, and Backpack, in Fig. 8. For the Backpack, those samples whose backpack strap color are similar to upper body clothes or putting the package in the front are hard to be recognized.

3) Influence of Occlusion: For a quantitative analysis on the influence of occlusion, we use clean data for training the SVM classifiers with Pool5 features and only the occlusion data for test. The recognition results of partial attributes are shown in Fig. 10. Compared with average results on clean test set, almost all the attributes' accuracies drop drastically, such as Female and Dress. The average accuracy of all the attributes drops 5.42%. However, the results of attributes like BodyFat which reflects the character from full body, change a little. This indicates that the attributes corresponding with body parts may be easier affected by occlusion, while the attributes of the full body are more robust to the partial occlusions.

Besides the influence of occlusion regions, we also make a quantitative analysis on the influence of occlusion type "person occluded by other people" on attribute recognition. Specifically, we train the attribute recognition models based on the cleaning training set without any occlusions and evaluate the models on two types of test sets, i.e., one is a clean test set without occlusions, the other one only includes the person samples occluded by other persons. The experimental results based on the metric of mA are 74.60% and 67.66%, respectively. Compared to the mean attribute recognition results on test set with all kinds of occlusions (69.18%), the occlusion type of person occluded by other person has lower performance, which shows that the person occlusion is still an important challenge in pedestrian attribute recognition.

Some challenging effects of "person occluded by other people" on pedestrian attribute recognition are shown in Fig. 9. As shown in Fig. 9, when multiple persons gather and serious occlusions occur, the model may easily shift from one person to another, where the recognized local attributes may be from different persons. For example, for the first person in Fig. 9, the model shifts from the man to the woman, which leads to the mistakes on gender, hair style and lower clothing. For the second person, the model shifts from the female in the middle to the male in the back and the female in

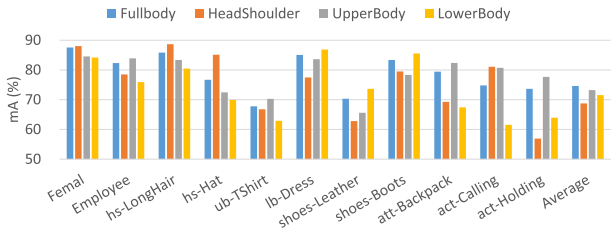


Fig. 11. Partial attributes' results in mA (%) at different parts. FullBody means the recognition results using the body image without parts. Best viewed in color.

Fig. 12. Visualization of partial attributes' results at three parts. The green boxes are true positives and the red boxes are false negatives. Best viewed in color.

the front. Thus, the person occlusions in the real scenes pose a challenging problem on person attribute recognition.

4) Influence of Body Parts: To analyze the influence of body parts, only the images without occlusions are selected in training and test stage. For each selected human image, the predefined three parts including head-shoulder, upper-body, lower-body are extracted based on ground-truth annotations. For each part, a linear SVM model with Pool5 features is trained for each selected attribute. In the test stage, the ground truth parts are used to recognize attributes, where the influence of part detection could be removed. Results of selected attributes are shown in Fig. 11.

For human body, some attributes are relevant to fixed parts. For example, Longhair depends on the HeadShoulder, and Tshirt exists on the UpperBody. This is consistent with our results in Fig. 11. For example, LongHair achieves high mean accuracy using HeadShoulder features. Besides the part-based attributes, some attributes, e.g. gender, are related to the FullBody, and they are not sensitive to body parts. This is also consistent with human perception. To have a better visualization, we show the recognition results in Fig. 12. Generally, those attributes which are partially occluded or have similar appearances with background are hard to recognize, such as LongHair and Hashat in HeadShoulder. There is still a long way to learn to recognize these hard samples.

In summary, recognizing attributes in UpperBody is easier than the other two body parts. This is reasonable as rich information exists in UpperBody. The recognition of part related attributes according to parts is relatively easier than just using full body information. Spatial knowledge guided pedestrian attribute recognition may be a promising direction.

TABLE X

ATTRIBUTE RECOGNITION RESULTS (%). THE BEST RESULTS EXCEPT THE LAST TWO ROWS ARE INDICATED IN **BOLD**. IACN AND IDEEPMAR ARE THE INDIVIDUAL ATTRIBUTE VERSIONS OF ACN AND DEEPMAR. "C" AND "R" MEANS CAFFENET AND RESNET50, RESPECTIVELY. DEEPMAR* ARE RESULTS BASED ON BODY PARTS

Methods	label-based	instance-based			
	mA	Accuracy	Precision	Recall	F1
SVM-ELF	65.60	19.70	21.54	67.39	32.65
SVM-Fc6-C	72.62	30.31	34.11	71.12	46.10
SVM-Fc7-C	71.81	29.67	33.38	70.76	45.36
SVM-Pool5-R	74.52	33.08	37.14	73.42	49.33
IACN-C [31]	66.63	60.79	**80.12**	70.15	74.81
ACN-C [31]	68.92	61.96	70.89	**80.90**	75.56
IDeepMAR-C [27]	72.29	60.22	73.15	75.83	74.47
DeepMAR-C [27]	72.94	61.13	74.21	75.22	74.71
DeepMAR-R [27]	**75.54**	**64.84**	76.56	78.64	**77.59**
DeepMAR*-C	76.01	63.78	76.13	78.12	77.11
DeepMAR*-R	75.54	65.60	80.06	76.54	78.26

5) Relationships Among Attributes: Some attributes have strong correlations with each other. Learning multiple attributes simultaneously may better capture the relationships among attributes than learning each attribute separately. The overall results are shown in Table X. As shown in Table X, both ACN and DeepMAR can obtain better results than their individual attribute versions IACN and IDeepMAR, as well as the corresponding SVM-based methods, on *label-based* evaluation and F1 of *instance-based* evaluation. This shows that learning multiple attributes jointly could better explore the relationships among attributes and improve the final results.

To further study the body parts' influence on attribute recognition, we perform extra experiments by utilizing the full body and body parts together to recognize attributes, which is shown as DeepMAR* in Table X. The input of DeepMAR* has four blocks, including body image, head-shoulder image, upper body image and lower body image. Then the features are concatenated together at the final feature layer ("FC7" or "Pool5"). The ground truth parts are used in the test stage. After utilizing the body and parts together, both *label-based* and *instance-based* scores can be improved by 3% in CaffeNet. Though the improvement is not obvious with ResNet50 as CaffeNet, it is still validate the advantage of fusing parts and global information for attribute recognition.

B. Attribute-Based Person Retrieval

Typical attribute-based person retrieval depends on the attribute recognition results. We adopt the same attribute recognition algorithms in Section IV-A, including SVM, ACN, and DeepMAR. In test stage, we use various numbers of attributes for the query, i.e., single attribute based query and multi-attribute based query. For the first category, we use all the 54 attributes in our experiments. For the second category, we manually select 5 group of attributes to generate various combinations as queries, including gender, 4 upper body attributes, 9 lower body attributes, 6 shoe types and 7 attachments. Not all the combinations of attributes are used because many combinations have less true positive samples. We then adopt a two-step strategy to sample the multiple attribute queries. First, one or more groups are randomly

Fig. 13. Partial results of multiple attributes based retrieval. The green boxes satisfy the query while the red ones do not. Best viewed in color.

TABLE XI

THE MAP (%) OF PERSON RETRIEVAL RESULTS USING MULTIPLE ATTRIBUTES ON RAP DATASET. ONE TO FOUR REPRESENTS THE NUMBER OF ATTRIBUTES USED FOR RETRIEVAL. IACN AND IDEEP-MAR ARE THE INDIVIDUAL ATTRIBUTE VERSIONS OF ACN AND DEEPMAR. "C" AND "R" MEANS CAFFENET AND RESNET50, RESPECTIVELY

Methods	Mutiple queries			
	One	Two	Three	Four
SVM-ELF	20.49	11.64	4.37	1.90
SVM-Fc6-C	24.15	16.51	6.40	2.92
SVM-Fc7-C	23.96	15.54	5.08	2.53
SVM-Pool5-R	25.85	16.36	5.38	1.73
IACN-C [31]	52.34	51.81	39.57	29.58
ACN-C [31]	55.65	55.59	42.65	31.19
IDeepMAR-C [27]	53.91	53.67	40.79	30.21
DeepMAR-C [27]	56.06	57.02	43.37	32.25
DeepMAR-R [27]	61.88	64.36	50.85	39.37

selected except for gender. Second, one attribute is randomly selected for each selected group. We combine the gender and the selected attributes, and select the combinations which have more than 100 ground truth samples as the final multi-attribute based queries. Totally, there are 150 multi-attribute based queries which are listed in supplementary materials. The experimental results with different numbers of queried attributes are presented in Table XI.

For the single attribute based person retrieval, although SVM and deep models have similar *mA* values in Tabel X, the *mAP* of SVM drops a lot compared with deep models as shown in Table XI. This may because the *mA* concerns more about the recall rate while the *mAP* pays more attention to the precision. With the increase of the number of attributes used for retrieval, the performance of both SVM and deep models decreases monotonically. This is reasonable that the recognition of multiple attributes simultaneously become harder with the increase of attributes. So the consistent recognition of multiple attributes in the same person is very important for multiple attributes based person retrieval. This is why we introduce the *instance-based* evaluation metrics where the accuracy of co-occurrence of multiple attributes in the same person can be reflected. To be more clear, some results of multiple attributes based person retrieval are shown in Fig. 13.

TABLE XII

PERSON REID RESULTS (%) ON RAP DATASET. "×" MEANS THE DISTRACTORS ARE NOT INCLUDED IN GALLERY SET AND "✓" MEANS THE DISTRACTORS ARE USED IN GALLERY SET. "C" AND "R" MEANS CAFFENET AND RESNET50, RESPECTIVELY

		L2		KISS [41]		XQDA [42]	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ELF [37]	×	1.82	0.92	4.78	2.53	2.53	2.01
	✓	1.18	0.53	2.82	1.62	1.64	1.21
LOMO [42]	×	5.32	2.33	2.57	1.15	7.26	4.53
	✓	3.61	1.50	2.14	0.78	5.17	2.93
GOG [39]	×	9.44	4.89	6.93	2.65	31.49	19.06
	✓	7.00	3.40	5.18	1.78	24.91	14.65
JSTL [44]	×	26.73	11.59	28.81	13.24	29.94	14.80
	✓	21.08	8.47	22.84	9.91	23.37	10.89
ResNet 101 [32]	×	63.84	41.25	62.40	39.09	61.77	39.78
	✓	54.50	33.36	52.75	31.71	52.19	31.89
ResNet 152 [32]	×	66.93	43.44	64.75	40.10	65.40	42.01
	✓	57.80	35.41	56.21	32.85	56.16	34.00
DenseNet 121 [52]	×	62.64	43.07	62.73	41.68	62.40	42.39
	✓	53.24	34.96	53.21	33.85	52.75	34.07
PAR [18]	×	59.76	38.56	59.04	36.90	57.73	37.74
	✓	50.60	30.87	50.15	29.76	48.56	30.02
MSCAN [38]	×	48.17	29.29	45.79	27.28	47.81	29.21
	✓	37.80	22.21	35.43	20.41	37.46	22.11
MuDeep [53]	×	45.97	26.26	38.13	18.96	44.60	25.57
	✓	37.20	20.17	30.30	14.53	36.06	19.67
HACNN [54]	×	70.69	47.96	68.48	44.33	70.20	46.96
	✓	62.09	40.21	60.52	37.46	61.77	39.41
ATT-C	×	15.84	6.39	11.97	3.77	13.55	5.99
	✓	12.02	4.38	9.14	2.69	9.54	3.93
IDE-C	×	42.32	24.11	42.11	21.66	45.42	27.32
	✓	33.85	18.46	34.67	17.23	37.20	21.20
IDE-ATT-C	×	45.27	26.26	42.07	20.39	48.06	28.43
	✓	36.85	20.19	35.62	16.70	39.54	22.18
ATT-R	×	26.66	11.67	33.56	15.19	19.74	9.61
	✓	20.90	8.34	27.53	11.52	14.69	6.52
IDE-R	×	58.61	37.03	58.15	35.78	57.19	36.43
	✓	49.61	29.51	49.29	28.77	48.17	28.88
IDE-ATT-R	×	59.25	38.21	58.34	36.21	57.72	37.48
	✓	50.44	30.77	49.94	29.45	48.76	29.93

TABLE XIII

DATA PARTITION ON PERSON RE-IDENTIFICATION

Data partition	Train	Test	Distractor
#Identities	1,295	1,294	-
#Images	13,178	13,460	14,947

C. Image-Based Person Retrieval

For the ReID task, we adopt the fixed data split as shown in Table XIII. Specifically, there are 300 and 298 identities who appear more than one day in the training and test set, respectively. We randomly sample one image in each day under each camera for each test person identity to form the query set, and there are totally 7,202 queries generated in the test stage. The proposed distractors are also used to evaluate the effectiveness of different algorithms on large-scale gallery set. We randomly select one sample of each identity for validation to determine the proper hyperparameters, such as rate of dropout, the weight λ in Eq. (7). Then we use all the person images in the training set for training. Specifically, for KISS, the features are preprocessed with PCA where 95% energy is kept. The overall results are shown in Table XII and visualization of partial results are shown in Fig. 14.

For hand-crafted features, the GOG shows better performance than the ELF and the LOMO with three distance

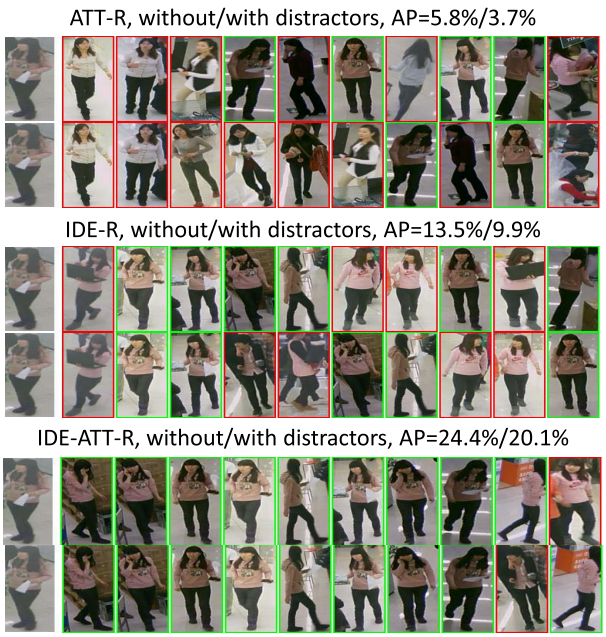


Fig. 14. Visualization of person retrieval with three methods. For each method, the first column is the query person and the rest in each row are gallery persons. Green and red boxes represent true positive and false positive samples, respectively. The first row and second row in each method are the rank list without and with distractors, respectively. Best viewed in color.

metrics. Compared with the hand-crafted features, the JSTL which is learned from multiple ReID datasets, shows better generalization capability. For those deep learning based methods, the end-to-end deep models, which have been trained on the RAP dataset, generally have obtained better performances than hand-crafted features ELF, LOMO, and GOG, as well as the pretrained deep features in JSTL. Among these deep models, the multi-scale networks [38], [53], which are specially designed to handle variation of pedestrian scale and pose, have achieved improved results. Meanwhile, the deeper networks, which own better representation ability, can obtain superior performance, such as ResNet152 [32] *vs.* ResNet101 [32], and DenseNet121 [52] *vs.* ResNet50 [32]. Furthermore, as a fine-grained zero-shot learning problem, person re-identification often resort to multi-branch networks to explore various visual cues from both the global and local information. The global branch could learn full body based features, while the local branch could learn body parts based features which can handle the pose variance and detection errors, as well as enhancing local discriminative ability. Based on this idea, the HACNN [54], which utilizes a multi-branch network to joint learn multiple complementary attention maps and maximizes their latent complementary effects, so as to achieve the state-of-the-art performance

For two backbone networks CaffeNet and ResNet50, the features which are learned from attribute classification (ATT-C and ATT-R) show weaker performance than identity classification (IDE-C and IDE-R). However, joint learning attribute and identity (IDE-ATT-C and IDE-ATT-R) can further improve person ReID results. This shows that attributes can indeed assist feature learning for person ReID. Compared with CaffeNet, the deeper network ResNet50 can produce better

TABLE XIV

ATTRIBUTE RECOGNITION RESULTS (%) ON THE TEST SET WITH/ WITHOUT DISTRACTORS. "×" MEANS THE DISTRACTORS ARE NOT INCLUDED IN GALLERY SET AND "✓" MEANS THE DISTRACTORS ARE USED IN GALLERY SET. "C" AND "R" MEANS CAFFENET AND RESNET50, RESPECTIVELY

Method		mA	Accuracy	Precision	Recall	F1
ATT-C	×	65.07	54.16	69.58	68.39	68.98
	✓	65.34	54.36	69.72	68.57	69.14
IDE-ATT-C	×	64.83	54.64	69.68	69.64	69.66
	✓	64.94	54.71	69.71	69.83	69.77
ATT-R	×	68.67	62.10	78.20	72.99	75.50
	✓	68.99	62.18	78.21	73.07	75.55
IDE-ATT-R	×	63.76	56.54	74.12	68.80	71.37
	✓	63.74	56.55	74.02	68.92	71.38

results, which is useful to guide the network design for ReID. Compared with different metric learning algorithms, we find that the L2 metric is good enough, and the learned KISS and XQDA have similar results or even little worse. This may because the learned feature metric can be more easily overfitting on the sampled pair-wise training data.

To explore the influence of ReID on attribute recognition, we show the attribute recognition results in Table XIV. For CaffeNet, the joint learning of attribute and identity has a little influence on attribute recognition, i.e., a little improvement based on the F1 metric, while a tiny decline for mA value. However, for ResNet50 which has much better representation ability, attribute recognition drops about 5% for mA value and about 4% for F1 metric. This is not consistent with results of Lin *et al.* [18], where both the ReID and attribute recognition can be improved. This may because that the ReID task can be seen as an instance-level person retrieval, while the attribute-based person retrieval is category-level. Thus, the objective function of ReID may weaken the invariance of attribute features. For example, many different persons may have some of the same attributes. Due to the learning of ReID features, more discriminative features should be learned, which leads to a larger variance in intra-attribute category.

The results [18] suggest that ReID can improve the results of attribute recognition. This is probably because the identity-based attribute annotation decreases the inconsistency between ReID and attribute recognition. However, the identity-based attribute annotation is not reasonable, as the visibility of attributes of the same person may vary along with environmental factors, e.g., viewpoints. Assume that the same person doesn't change the appearance as crossing multiple cameras, the identity-based attribute annotation subjects an exactly identical attribute based representation among intra-identity samples, which is consistent with the objective of ReID to learn an invariant representation for each identity. However, this assumption cannot well hold in RAP since the same person may have large variation due to the large interval of occurrence times (e.g., across days), and sometimes his/her clothing and appearance could change greatly over multiple days. So attribute classification may perform worse in the multi-task network in RAP.

Due to the long-term data collection, person identities may change their clothing and appearance greatly. To explore the

TABLE XV

REID RESULTS (%) IN ONE DAY ("I") AND CROSS DAY ("X"). "C" AND "R" MEANS CAFFENET AND RESNET50, RESPECTIVELY

		L2		KISS		XQDA	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IDE-ATT-C	I	59.14	46.37	51.93	35.51	61.97	50.29
	X	32.10	27.91	27.65	21.11	33.73	30.53
IDE-ATT-R	I	72.03	60.67	71.21	58.20	71.59	60.48
	X	41.42	38.28	41.43	36.32	41.10	38.01

TABLE XVI

DISTRIBUTIONS OF THE NUMBER OF QUERIES AND PERSON IDENTITIES AT DIFFERENT TYPES OF APPEARANCE CHANGES

	None	Upper	Lower	All
#Queries	1506	703	311	168
#Identities	251	116	76	43

influence of person retrieval in cross day, we select 3,397 query images whose identities appear over multiple days for intra-day and cross-day person retrieval. The overall results are shown in Table XV. Obviously, it is much harder to retrieve persons in the condition of cross day, which poses a new challenge for the research of person retrieval in the future. Furthermore, we make a quantitative analysis on the influence of appearance change on cross-day ReID. First, we partition the appearance change into four types, including less appearance change (None), only larger upper body appearance change (Upper), only larger lower body appearance change (Lower), larger upper and lower body appearance change (All). The number of queries and identities at different types of appearance changes are shown in Table XVI. Since some person identities may appear more than two days, so they may have more than one type of appearance change, which leads to the total number of identities over different types of appearance changes is 486, rather than 298. The appearance at corresponding region is defined as larger change if the corresponding cloth types have changed more than 50% or color types have changed more than 25%. Then based on these four types of appearance changes, we can test the cross-day ReID results at these four conditions independently with two backbone networks, e.g. CaffeNet and ResNet50. Finally, the quantitative ReID results under four types of appearance change are shown in Table XVII. The experimental results have shown that the mAP will decrease along the order of "None, Lower, Upper, All". It shows that the upper body has more impact than the lower body in ReID, which is consistent with common sense where the upper clothes often have more variations. Specifically, from "None" to "All", the mAP has decreased greatly, which means that the current person re-identification algorithms still have a large potential to handle the large appearance variance in long term person retrieval. Visualization of ReID results on the single day and cross day based on IDE-ATT-R under L2 metric are shown in Fig. 15.

Finally, we explore the influences of *instance-based* and *identity-based* annotations on attribute recognition and ReID. In RAP, we can transform instance-based attribute annotations into identity-based annotations, where for each identity, the attribute is annotated as positive if any of his images

Fig. 15. Visualization of person retrieval on one day (top) and cross day (bottom) for the same query. Green box and red box represents true positive and false positive samples, respectively. The same person has different upper clothes in two days. Best viewed in color.

TABLE XVII

QUANTIZED RESULTS (%) OF CROSS-DAY PERSON REID WITH DIFFERENT TYPES OF APPEARANCE CHANGE. "C" AND "R" MEANS CAFFENET AND RESNET50, RESPECTIVELY

		L2		KISS		XQDA	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IDE-ATT-C	None	36.07	31.33	31.19	23.94	37.76	34.13
	Upper	11.35	10.61	10.25	7.69	13.27	12.55
	Lower	27.87	22.94	21.64	13.98	29.51	25.12
	All	10.65	8.28	5.48	4.55	7.74	8.32
	Mean	32.10	27.91	27.65	21.11	33.73	30.53
IDE-ATT-R	None	46.06	42.44	45.80	40.31	45.45	42.13
	Upper	15.42	15.79	16.16	14.46	16.11	15.63
	Lower	44.10	38.26	48.20	37.70	46.56	38.56
	All	13.55	12.57	12.58	10.58	13.87	12.32
	Mean	41.42	38.28	41.43	36.32	41.10	38.01

TABLE XVIII

COMPARISON OF INSTANCE-BASED (PI) AND IDENTITY-BASED (PID) ANNOTATIONS ON THE RESULTS (%) OF REID AND ATTRIBUTE RECOGNITION. "C" AND "R" MEANS CAFFENET AND RESNET50, RESPECTIVELY

	Train	ReID		Test	Attribute recognition				
		Rank-1	mAP		mA	Acc.	Pre.	Rec.	F1
IDE-ATT-C	PI	36.85	20.19	PI	64.94	54.71	69.71	69.83	69.77
				PID	59.89	47.56	56.67	79.39	66.14
	PID	39.61	22.49	PI	64.32	42.95	47.73	83.62	60.77
				PID	62.00	48.70	75.10	62.78	68.39
IDE-ATT-R	PI	50.44	30.77	PI	63.74	56.55	74.02	68.92	71.38
				PID	59.32	48.21	55.49	83.23	66.58
	PID	48.76	29.61	PI	64.20	44.02	48.70	84.94	61.90
				PID	61.92	50.17	76.35	64.37	69.85

have the attribute. Based on the transformed annotations, we train a multi-task network as described in Section IV-C, and then evaluate attribute recognition and ReID on the test set with distractors. Experimental results of ReID and attribute recognition based on different combinations of two types of attribute annotations at training and test stages are shown in Table XVIII.

For attribute recognition, the results drop significantly with the instance-level metrics (e.g., F1) from PI to PID under instance-based test set for both CaffeNet and ResNet50, which illustrates that the identity-based attribute annotation makes the learned model confused to obtain a consistent attribute representation and weaken the accuracy of the prediction of multiple attributes on one image sample. Furthermore, to analyze the relations between the results of attribute recognition based on the two types of annotations in test set and those of person re-identification, we show the results in Fig. 16 to have a clear illustration. From the figure,

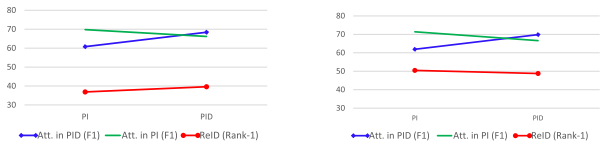


Fig. 16. The relations between the changing trends of ReID with Rank-1 metric (%) and the attribute recognition with F1 metric (%) with different annotation strategies. "C" and "R" means CaffeNet and ResNet50, respectively. (a) IDE-ATT-C. (b) IDE-ATT-R.

we can find the consistency between the changing trends of attribute and ReID is different depending on different learning models. When using the model CaffeNet (IDE-ATT-C), the PID based annotation in the test set results in a consistent changing trend of results of attribute recognition with those of ReID. As with deeper ResNet50 (IDE-ATT-R), the attribute recognition results based on PI based annotations in the test set are more consistent with the ReID performance. While for the ReID task, as shown in [18], the identity-based attribute annotation subjects an exactly identical attribute-based representation among intra-identity samples, so that the ReID results can be improved form PI to PID under CaffeNet. However, for the network ResNet50 which owns more powerful capability to learn intrinsic image features, the more accurate instance-based annotations obtain superior performance than the identity-based annotations. Considering the somewhat inconsistent objective of attribute recognition and ReID, it is worthy to study the problem of attribute-assisted ReID with new designed objective function or network structures in future.

In summary, with the RAP dataset, researchers can explore more possibilities to improve person-centric perception algorithms based on the richly annotated person attributes and IDs. Firstly, the generative adversarial networks (GAN) based domain adaption has attracted great attentions in ReID researches recently, such as Similarity Preserving Generative Adversarial Network (SPGAN) [56] and Person Transfer Generative Adversarial Network (PTGAN) [57]. Based on the RAP dataset, the GAN can be improved to synthetize more high-fidelity pedestrian images with some guidance of attribute consistency. Secondly, towards a middle-level attribute based representation, the interpretability of ReID model can also be enhanced, where the fine-grained similarity at attribute level can be studied to provide more reliable explanations of person ReID model. Finally, the RAP dataset will also extended with more data and annotations in the future, such as person poses, and dynamic tracklets, so that the multiple kinds of rich annotations can provide a new evaluation platform of composable intelligent vision system.

VI. Conclusions

To promote the research of person retrieval, we collected a large-scale richly annotated pedestrian dataset in real surveillance scenarios. Based on the dataset, we implement several state-of-the-art algorithms on pedestrian attribute recognition and person re-identification, and perform extensive evaluations with three tasks on person retrieval, where an effective instance-based metric is proposed to measure the performance of attribute recognition. We also have made detailed analyses about the contextual and environmental factors in pedestrian attribute recognition. Furthermore, we explore the influence of pedestrian attribute for person re-identification based on current state-of-the-art deep learning networks with a multi-task learning structure. In addition, a new challenging problem of cross-day person retrieval is posed and we find that the current re-identification algorithms are still not good enough to handle the partial variation of cloth appearance. To our knowledge, the RAP dataset is current largest person retrieval dataset which can support attribute-based person retrieval and person ReID simultaneously. The rich annotations of attributes, person identities as well as the environmental and contextual factors provides a unified testbed to develop more advanced techniques on person retrieval in real surveillance scenarios.

Acknowledgment

The authors would like to thank all the anonymous reviewers and the associate editor for their valuable comments to improve the paper. The authors would also like to thank Haibin Ling, Da Li, Yang Zhou, Weihua Chen, Houjing Huang, Wenjie Yang, for their helps in improving the paper, as well as the dataset.

References

[1] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti, "Attribute-based people search: Lessons learnt from a practical surveillance system," in *Proc. Int. Conf. Multimedia Retr.*, 2014, p. 153.

[2] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1521–1528.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann. (Oct. 2016). "Person re-identification: Past, present and future." [Online]. Available: https://arxiv.org/abs/1610.02984

[4] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5098–5107.

[5] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 789–792.

[6] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Pedestrian attribute classification in surveillance: Database and evaluation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 331–338.

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

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

[9] Z. Zheng, L. Zheng, and Y. Yang. (2017). "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*." [Online]. Available: https://arxiv.org/abs/1701.07717

[10] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, vol. 3, no. 5, Oct. 2007, pp. 1–7.

[11] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. Berlin, Germany: Springer, 2011, pp. 91–102.

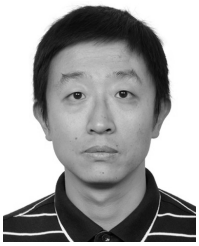
[12] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2012, pp. 391–401.

[13] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[14] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Proc. Workshop Appl. Comput. Vis. Workshops (WACV)*, Dec. 2009, pp. 1–8.

[15] C. Luo, Z. Li, K. Huang, J. Feng, and M. Wang, "Zero-shot learning via attribute regression and class prototype rectification," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 637–648, Feb. 2018.
[16] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5187–5196.
[17] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, vol. 2, no. 3, 2012, p. 8.
[18] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. (2017). "Improving person re-identification by attribute and identity learning." [Online]. Available: https://arxiv.org/abs/1703.07220
[19] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, no. 6, pp. 23.1–23.11, 2009.
[20] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asia Conf. Comput. Vis.*, 2012, pp. 31–44.
[21] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3594–3601.
[22] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 613–627.
[23] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2012, pp. 402–412.
[24] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.
[25] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[27] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 111–115.
[28] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1637–1644.
[29] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1543–1550.
[30] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label CNN based pedestrian attribute learning for soft biometrics," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 535–540.
[31] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2015, pp. 87–95.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[33] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen, "Person attribute search for large-area video surveillance," in *Proc. IEEE Int. Conf. Technol. Homeland Secur. (HST)*, Nov. 2011, pp. 55–61.
[34] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4184–4193.
[35] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, "Soft biometric retrieval to describe and identify surveillance images," in *Proc. IEEE Int. Conf. Identity, Secur. Behav. Anal. (ISBA)*, Feb./Mar. 2016, pp. 1–6.
[36] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*, vol. 1. Berlin, Germany: Springer, 2014.
[37] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.
[38] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 384–393.
[39] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1363–1372.
[40] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, and O. Camps, "Person re-identification in appearance impaired scenarios," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 48.1–48.14.
[41] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.
[42] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
[43] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 403–412.
[44] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1249–1258.
[45] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1367–1376.
[46] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 475–491.
[47] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. IEEE Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2428–2433.
[48] T. Nortcliffe, "People analysis cctv investigator handbook," *Home Office Centre Appl. Sci. Technol.*, vol. 2, no. 3, pp. 1–35, 2011.
[49] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[50] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[51] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 801–808.
[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.
[53] X. Qian, Y. Fu, Y.-G. Jiang, and T. X. X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5399–5408.
[54] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.
[56] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.
[57] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

Dangwei Li received the B.S. degree from Jilin University, Changchun, China, in 2013. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is also with the University of Chinese Academy of Sciences. His current research interests are computer vision, deep learning, and pedestrian analysis.



Zhang Zhang received the B.S. degree in computer science and technology from the Hebei University of Technology, Tianjin, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an Associate Professor at the Center for Research on Intelligent Perception and Computing and the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He has published a number of papers at top venues, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS and *Machine Intelligence*, CVPR, and ECCV. His research interests include activity recognition, video surveillance, and time series analysis.

Xiaotang Chen (M'16) received the B.E. degree from Xidian University in 2008 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2013. In 2013, she joined CASIA as an Assistant Professor. Her major research interests include computer vision and pattern recognition. She served as the technical program committee member of several conferences.

Kaiqi Huang (SM'08) received the B.Sc. and M.Sc. degrees from the Nanjing University of Science Technology, China, and the Ph.D. degree from Southeast University. He has been a Full Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, since 2005. He has published over 180 papers in the important international journals and conferences, such as the IEEE TPAMI, T-IP, T-SMCB, TCSVT, *Pattern Recognition*, CVIU, ICCV, ECCV, CVPR, ICIP, and ICPR. His current researches focus on computer vision and pattern recognition, including object recognition, video analysis, and visual surveillance. He is the Deputy General Secretary of the IEEE Beijing Section from 2006 to 2008. He is an Executive Team Member of the IEEE SMC Cognitive Computing Committee. He received the winner prizes of the object detection tasks in both PASCAL VOC'10 and PASCAL VOC'11, object classification task in ImageNet-ILSVRC2014. He received some awards, including The National Science and Technology Progress Award in 2011, the Excellent Young Scholars Award of National Science Foundation of China in 2013, and the CCF-IEEE Young Scientist Award in 2016. He serves as co-chairs and program committee members over 40 international conferences, such as ICCV, CVPR, ECCV, and the IEEE workshops on visual surveillance. He is an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS and *Pattern Recognition*.