Pattern Recognition 75 (2018) 77–89

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-type attributes driven multi-camera person re-identification

Chi Su[a], Shiliang Zhang[a,*], Junliang Xing[b], Wen Gao[a], Qi Tian[c]

[a] National Engineering Laboratory for Video Technology, Peking University, Beijing, China
[b] National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China
[c] Department of Computer Science, University of Texas at San Antonio, San Antonio, USA

ARTICLE INFO

Article history:
Received 25 September 2016
Revised 1 June 2017
Accepted 4 July 2017
Available online 12 September 2017

Keywords:
Deep attributes
Person re-identification

ABSTRACT

One of the major challenges in person Re-Identification (ReID) is the inconsistent visual appearance of a person. Current works on visual feature and distance metric learning have achieved significant achievements, but still suffer from the limited robustness to pose variations, viewpoint changes, *etc*., and the high computational complexity. This makes person ReID among multiple cameras still challenging. This work is motivated to learn mid-level human attributes which are robust to visual appearance variations and could be used as efficient features for person matching. We propose a weakly supervised multi-type attribute learning framework which considers the contextual cues among attributes and progressively boosts the accuracy of attributes only using a limited number of labeled data. Specifically, this framework involves a three-stage training. A deep Convolutional Neural Network (dCNN) is first trained on an independent dataset labeled with attributes. Then it is fine-tuned on another dataset only labeled with person IDs using our defined triplet loss. Finally, the updated dCNN predicts attribute labels for the target dataset, which is combined with the independent dataset for the final round of fine-tuning. The predicted attributes, namely *deep attributes* exhibit promising generalization ability across different datasets. By directly using the deep attributes with simple Cosine distance, we have obtained competitive accuracy on four person ReID datasets. Experiments also show that a simple distance metric learning modular further boosts our method, making it outperform many recent works.

1. Introduction

Person Re-Identification (ReID) is a technology to identify the same person across images captured by different cameras. As is shown in Fig. 1, person ReID is challenging because the visual appearance of a person is easily affected by many factors, including illumination conditions, viewpoint variations, camera parameters, body poses, *etc*. Due to its important applications in public security, *e.g.*, cross camera pedestrian searching, tracking, and event detection, person ReID has attracted lots of attention from both the academic and industrial communities. Currently, most research efforts can be summarized into two categories: a) extracting and encoding robust local features representing the visual appearance of a person [1–7] and b) reducing the distance between features of the same person by learning a discriminative distance metric [8–25].

Despite the significant achievements made by existing works, there is still much room for improvement before person ReID can be used in real applications. Because local features mainly describe low-level visual appearance, they are not robust to variances of viewpoints, body poses, *etc*. On the other hand, distance metric learning suffers from the poor generalization ability and the quadratic computational complexity, *e.g.*, different datasets present different visual characteristics corresponding to different metrics. Compared with low-level visual feature, human attributes like long hair, blue shirt, *etc*., represent mid-level semantics of a person. As illustrated in Fig. 1, attributes are more consistent for the same person and are more robust to the above mentioned variances. Some recent works hence have started to use attributes for person ReID [26–31]. Because human attributes are expensive for manual annotation, it is difficult to acquire enough training data for a large set of attributes. This limits the performance of current attribute features. Consequently, low-level visual features still play a key role and attributes are mostly used as auxiliary features [28–31].

Recently, deep learning has exhibited promising performance and generalization ability in various vision tasks. For example in [35], an eight-layer deep Convolutional Neural Network (dCNN) is trained with large-scale images for visual classification. The modified versions of this network also perform impressively in object detection [36] and segmentation [37]. Many researchers have started to use deep learning and triplet loss for person ReID

* Corresponding author.
E-mail address: slzhang.jdl@pku.edu.cn (S. Zhang).

http://dx.doi.org/10.1016/j.patcog.2017.07.005
0031-3203/

Fig. 1. Example images of the same person taken by two cameras from three datasets: (a) *VIPeR* [32], (b) *PRID* [33], and (c) *GRID*[34]. This figure also shows five of our predicted attributes shared by these two images.

[38–40]. Specifically, they use two images of the same person and one image of another person to construct a triplet. Then, triplet loss backpropagated to update the dCNN to learn a discriminative distance metric, i.e., the distance between two images of the same person should be smaller.

Inspired by the promising performance of attributes and the strong generalization ability of dCNN, we target to learn a dCNN to detect a large set of human attributes for person ReID. Due to the diversity and complexity of human attributes, it is a laborious task to manually label enough of attributes for dCNN training. The key issue is hence how to train this dCNN from a partially-labeled dataset and ensure its discriminative power and generalization ability in the person ReID tasks. Meanwhile, some attributes are not compatible with each other. For example, gender-related attributes such "female" and male can not coexist for the same person. It is also not reasonable to predict multiple positive hair-related attributes like "hairLong", "hairBald", "hairShort", *etc.* for the same person. Therefore, instead of using the flat multi-label prediction structure, we should design a proper dCNN structure to take such contextual cues into consideration.

To address these issues, we propose a Weakly Supervised Multi-Type Attribute Learning (WSMTAL) algorithm. As shown in Fig. 2, we divide human attributes into multiple types, where each contains several incompatible attributes and only one of them can be positive. For example, the gender-related attributes and hair-related attributes would belong to two different types of attributes. In our dCNN, different types of attributes share the same convolutional layers, but each has its own fully connected layers and Softmax out layer to ensure the label incompatibility. Our WSMTAL is proposed to train this network with three stages.

In WSMTAL, the dCNN is firstly trained with the independent dataset, then is refined to acquire more discriminative power for person ReID task. Because this procedure involves one dataset with attribute labels and another without attribute labels, we call it a weakly supervised learning. Moreover, we divide the attributes into different types to ensure the incompatibility among attributes within each type. The attributes predicted by the final dCNN model are named as *deep attributes*. This structure is more reasonable than our previous work [41], which detects multiple attributes with a flat cross-entropy output layer.

To validate the performance of deep attributes, we test them on four popular person ReID datasets *without* combining with the local visual features. The experimental results show that deep attributes perform well, *e.g.*, they outperform many recent works combining both attributes and local features [28–31]. Note that, predicting and matching deep attributes make person ReID system faster, because it no longer needs to extract and code local features, compute distance metric, and fuse with other features.

Our contributions can be summarized as follows:

- We propose a three-stage weakly-supervised deep attribute learning algorithm, which makes learning a large set of human attributes from a limited number of labeled attribute data possible.
- An attribute triplet loss is proposed to predict attributes into multiple types and consider contextual cues among attributes.
- Deep attributes achieve promising performance and generalization ability on four person ReID datasets. Moreover, deep attributes release the previous dependencies on local features, thus have the potential to make the person ReID systems more robust and efficient.

This work extends our conference version [41] in the following aspects:

- Our original SSDAL directly learns attributes using plain sigmoid cross-entropy loss. The proposed WSMTAL model splits attributes into many types, where each includes several incompatible attributes and only one of them can be positive. This structure considers extra contextual cues among attributes and results in better performance.
- Our original SSDAL selects positive attributes by referring to thresholds, which vary on different datasets and are hard to decide. This shortcoming has been effectively addressed by WSMTAL. WSMTAL splits attributes into C types, where each includes several incompatible attributes and only one of them can be positive. In this way, C positive attributes can be identified in C classification tasks.
- More extensive experiments are conducted to test the validity of our approach. More comparisons to recent works, as well as the prediction accuracy of each type of attributes have been added. Deeper network, i.e., the VGG network is tested in our framework. This shows that our model is compatible with different deep networks, thus could leverage latest deep models to further improve the person Re-ID performance.

2. Related work

In this section, we briefly summarize and discuss related works in four aspects, i.e., 1) traditional low-level feature and distance metric learning based person ReID, 2) attributes based person ReID, 3) deep learning for attributes prediction, and 4) deep learning for person ReID, respectively.

Many researchers extract and encode low-level features for person ReID [1–7]. To handle viewpoint changes, Farenzena et al. [1] devise the Symmetry-Driven Accumulation of Local Features (SDALF) by the symmetric nature of pedestrians appearance. Cheng et al. [2] use pictorial structures and compute visual features in different parts of the body to estimate human body configuration to tackle the pose variations issue. There are also may methods measure the similarity between images of two different cameras [8–23,25] by learning a more reasonable distance metric. The Relaxed Pairwise Metric Learning (RPML) [10] is a method of relaxing the original hard constraints so as to make the computation more efficient. Zheng et al. [17] also introduce a Probabilistic Relative Distance Comparison (PRDC) model. Shen et al. [22] learn a correspondence structure using boosting which represents the two images features from a target camera pair matching probabilities. Considering the positive semi-definite constraint, Liao et al. [23] use a logistic metric learning approach to perform person ReID.

Attributes are efficient and discriminative for person ReID and have been used as features in may works [26–31]. Layne et al.

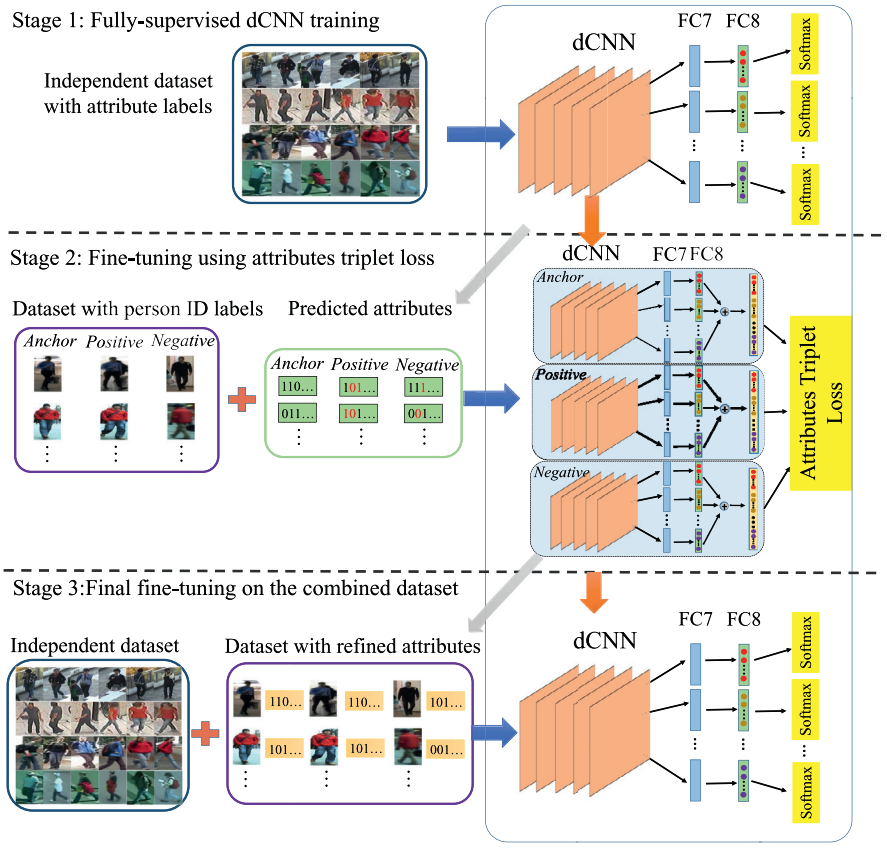

Fig. 2. Illustration of our dCNN structure and the Weakly Supervised Multi-Type Attribute Learning (WSMTAL) algorithm.

[28] show attributes improve ReID accuracy when combined with low-level features. Su et al. [31] present a low rank attribute embedding framework for person ReID using a novel multi-task learning framework. However, most of these methods use attributes as auxiliary information to aid the low level features.

Currently, many studies have applied deep learning to attributes learning [42–48]. Shankar et al. [42] propose a deep-carving neural net to learn attributes for natural scene images. Chen et al. [43] use a double-path deep domain adaptation network to get the fine-grained clothing attributes. Li et al. [44] propose two deep learning models to learn the pedestrian attributes, one is called as deep learning based single attribute recognition model (DeepSAR) and the other is a deep learning framework which recognizes multiple attributes jointly (DeepMAR). Huang et al. [45] propose a Dual Attribute-aware Ranking Network (DARN) to represent deep features using attribute-guided learning for cross-domain image retrieval. Yu et al. [46] use their weakly supervised deep learning model not only to recognize attribute but also to exploit the locations and rough shapes of pedestrian attributes. Our work differs from them in the aspects of both motivation and methodology. We are motivated by how to learn attributes of the human cropped from surveillance videos from a small set of data labeled with attributes. Our weakly supervised learning framework consistently boosts the discriminative power of dCNN and attributes for person ReID.

The works by Zhu et al. [47,48] are earlier works using deep learning for attribute based person ReID. They use a multi-label convolutional neural network (MLCNN) to predict multiple attributes with body part division. Zhu et al. finally combine deep attributes and low-level features and get promising person re-identification performance. Different from their work, our algorithm is more concise and efficient, i.e., does not use body part division to learn attributes and directly uses deep attributes to perform person ReID.

Inspired by the promising performance of deep learning, some researchers begin to use deep learning to learn visual features and distance metrics for person ReID [24,38,39,49–51]. In [49], Li et al. use a deep filter pairing neural network for person ReID, where two paired filters of two cameras are used to automatically learn optimal features. In [50], Yi et al. present a “siamese” convolutional network for deep distance metric learning. In [51], Ahmed et al. devise a deep neural network structure to transform person re-identification into a problem of binary classification, which judges whether a pair of images from two cameras contain the same person. In [24], Ding et al. present a scalable distance learning framework based on the deep neural network with the triplet loss. Chen et al. [38] propose a novel multi-channel parts-based convolutional neural network model with the triplet loss for person ReID. They also use a new threshold to improve the triplet loss. Wang et al. [39] present a joint learning framework to unify single-image representation and classification of cross-image representation using dCNN.

Despite of their efforts to find better visual features and distance metrics, the above mentioned works are designed specifically for certain datasets and are dependent on their camera settings. Differently, we use deep learning to acquire general camera-

independent mid-level representations. As a result, our algorithm shows better flexibility, *e.g.*, it could handle person ReID tasks on datasets containing different number of cameras.

3. Proposed approach

3.1. Framework

Our goal is to learn a large set of human attributes for person ReID through dCNN training. We define $A = \{A^1, A^2, \ldots, A^C\}$ as the collection of K attributes belonging to C types, and $A^c = \{a_1^c, a_2^c, \ldots, a_{K^c}^c\}$ denotes the label of the cth type attribute containing K^c attributes, where $a \in \{0, 1\}$ is the binary label. We divide attributes in to C types, and ensure $\forall i \neq j, A^i \cap A^j = \emptyset$. Therefore, $A^c \subset A$, $\sum_{c=1}^{C} K^c = K$. Our goal is thus learning an attribute detector $\mathcal{O}$, which predicts the attribute labels A for any input image I, i.e.,

$$A_I = \mathcal{O}(I). \tag{1}$$

Because of the promising discriminative power and generalization ability, we use dCNN model as the detector $\mathcal{O}(\cdot)$. However, dCNN training requires large-scale training data labeled with human attributes. Manually collecting such data is also too expensive to conduct. To ensure effective learning of a dCNN model for person ReID from only a small amount of labeled training data, we propose the Weakly supervised Multi-Type Attribute Learning (WSMTAL) algorithm.

As shown in Fig. 2, in the first training stage, an independent dataset with attribute labels is used to perform fully-supervised dCNN training. The resulting dCNN produces initial attribute labels for the target dataset. To improve the discriminative power of these attributes for ReID task, we start the second stage of training, i.e., fine-tuning the network using the person ID labels and our defined *attributes triplet loss*. The attributes triplet loss updates the network to enforce that the same person has more similar attributes and vice versa. The training data for fine-tuning can be easily collected because the person ID labels are readily accessible in many person tracking datasets. This fine-tuned dCNN hence predicts updated attribute labels for target datasets. Finally in the third stage, the labeled target dataset plus the original independent dataset are combined for the final stage of fine-tuning. The attributes predicted by the final dCNN model are named as *deep attributes*.

3.2. Fully-supervised dCNN training

We define the independent training set T with their attribute labels as $A_T = \{A_T^1, A_T^2, \ldots, A_T^C\}$. In T, each sample is annotated with a binary attribute label, *e.g.*, the label of the nth instance T_n is $A_{T_n} = \{A_{T_n}^1, A_{T_n}^2, \ldots, A_{T_n}^C\}$.

In the first stage of training, we use T as the training set for fully-supervised learning. We refer to the 16-layer VGG network [52] to build our dCNN model for its promising performance in various vision tasks. Specifically, our dCNN is also a 16-layer network, including 13 convolutional layers and 3 fully connected layers, where the 3rd fully connected layer predicts the attribute labels. The kernel and filter sizes of each layer in our architecture are the same with the ones in [52].

Our dCNN is shown in Fig. 3. We suppose that each type of attributes can only has one positive prediction to ensure its label incompatibility. Therefore, it is natural to use Softmax layer, which outputs only one positive prediction, for each type of attributes. In this way, C types of attributes can be predicted for each image. We denote the dCNN model learned in this stage as $\mathcal{O}^{S1}$. $\mathcal{O}^{S1}$ could predict attribute labels for any test sample.

However, as illustrated in our experiments, the discriminative power of $\mathcal{O}^{S1}$ is weak because of the limited scale and label accuracy of the independent training set. Therefore, We proceed to introduce our second stage of training.

3.3. dCNN fine-tuning with attributes triplet loss

In the second stage, a larger dataset is used to fine tune the previous dCNN model $\mathcal{O}^{S1}$. The goal of our dCNN model is predicting attribute labels for person ReID tasks. The predicted attribute labels thus should be similar for the same person. Motivated by this, we use person ID labels to fine-tune $\mathcal{O}^{S1}$ and produce similar attribute labels for the same person and vice versa. We denote the dataset with person ID labels as $U = \{u_1, u_2, \ldots, u_M\}$, where M is the number of samples and each sample has a person ID label l, *e.g.*, the mth instance u_m has person ID l_m.

In the second stage of training, we first use $\mathcal{O}^{S1}$ to predict the attribute label $\tilde{A}$ of each sample in U. For each sample, we concatenate the outputs of C Softmax classifiers as the attribute label. Thus, for the attribute label $\tilde{A_m}$ of the mth sample, we get C positive attributes. Then, we use the person ID labels to measure the annotation errors of $\mathcal{O}^{S1}$.

The annotation error of the $\mathcal{O}^{S1}$ is computed among three samples. The three samples are randomly selected from the U through the following steps: 1) select an *anchor* sample $u_{(a)}$, 2) select another *positive* sample $u_{(p)}$ with the same person ID with $u_{(a)}$, and 3) select a *negative* sample u_n with different person ID. Thus, a triplet $[u_{(a)}, u_{(p)}, u_{(n)}]$ is constructed, where the subscripts (a), (p), and (n) denote *anchor*, *positive*, and *negative* samples, respectively. The attributes of the eth triplet predicted by $\mathcal{O}^{S1}$ are $\tilde{A}_{(a)}^{(e)}$, $\tilde{A}_{(p)}^{(e)}$, and $\tilde{A}_{(n)}^{(e)}$ at the beginning of the fine-tuning, respectively.

The objectives of the fine-tuning is minimizing the triplet loss through updating the $\mathcal{O}^{S1}$, i.e., minimize the distance between the attributes of $u_{(a)}$ and $u_{(p)}$, meanwhile maximize the distance between $u_{(a)}$ and $u_{(n)}$. We call this triplet loss as attributes triplet loss. We hence could formulate our objective function for fine-tuning as:

$$\mathbf{D}\left(A_{(a)}^{(e)}, A_{(p)}^{(e)}\right) + \theta < \mathbf{D}\left(A_{(a)}^{(e)}, A_{(n)}^{(e)}\right), \ \forall \left(A_{(a)}^{(e)}, A_{(p)}^{(e)}, A_{(n)}^{(e)}\right) \in \mathcal{T}, \tag{2}$$

where $\mathbf{D}(.)$ represents the distance function of the two binary attribute vectors, $A_{(a)}^{(e)}$, $A_{(p)}^{(e)}$ and $A_{(n)}^{(e)}$ are predicted attributes of the eth triplet during the fine-tuning. Then, the corresponding loss function can be formulated as:

$$\mathcal{L} = \sum_{e}^{E} \max\left(0, \mathbf{D}\left(A_{(a)}^{(e)}, A_{(p)}^{(e)}\right) + \theta - \mathbf{D}\left(A_{(a)}^{(e)}, A_{(n)}^{(e)}\right)\right), \tag{3}$$

where E represents the number of triplets. In Eq. (3), if the $\mathbf{D}\left(A_{(a)}^{(e)}, A_{(n)}^{(e)}\right) - \mathbf{D}\left(A_{(a)}^{(e)}, A_{(p)}^{(e)}\right)$ is larger than θ, the loss would be zero. Therefore, parameter θ largely controls the strictness of the loss.

The above loss function essentially enforces the dCNN to produce similar attributes for the same person. However, the person ID label is not strong enough to train the dCNN with accurate attributes. Without proper constraints, the above loss function may generate meaningless attribute labels and easily over-fit the training dataset U. For example, imposing a large number meaningless attributes to two samples of a person may decrease the distance between their attribute labels, but does not help to improve the discriminative power of the dCNN. Therefore, we add several regularization terms and modify the original loss function as:

$$\mathcal{L} = \sum_{e}^{E} \left\{ \max\left(0, \mathbf{D}\left(A_{(a)}^{(e)}, A_{(p)}^{(e)}\right) + \theta - \quad \mathbf{D}\left(A_{(a)}^{(e)}, A_{(n)}^{(e)}\right)\right) + \gamma \times \mathcal{E} \right\} \tag{4}$$

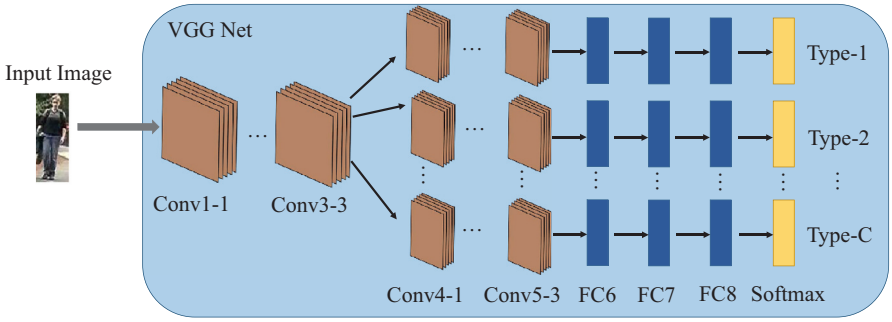

Fig. 3. The architecture of our dCNN. The network takes a RGB image as input. All types of attributes share Conv1-1 to Conv3-3 parameters and have their own independent parameters from Conv4-1 to FC8. Within each type of attributes, we use Softmax layer to ensure the label incompatibility.

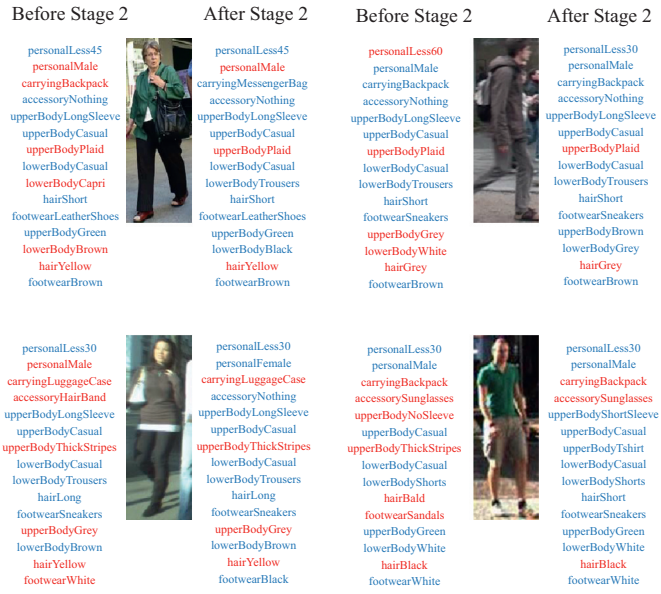

Fig. 4. Examples of predicted attributes on *MOT challenge* [53] by the learned dCNN before Stage 2 and after Stage 2. Texts with blue color are correct attributes, while those with red color are false attributes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\mathcal{E} = \mathbf{D}\left(A_{(a)}^{(e)}, \tilde{A}_{(a)}^{(e)}\right) + \mathbf{D}\left(A_{(p)}^{(e)}, \tilde{A}_{(p)}^{(e)}\right) + \mathbf{D}\left(A_{(n)}^{(e)}, \tilde{A}_{(n)}^{(e)}\right), \tag{5}$$

where $\mathcal{E}$ denotes the amount of change in attributes caused by the fine-tuning. The loss in Eq. (4) not only ensures that the same person has similar attributes, but also avoids the meaningless attributes. We hence use the above loss to update the $\mathcal{O}^{S1}$ with back propagation. We denote the resulting update dCNN as $\mathcal{O}^{S2}$.

In Fig. 4, we show examples of predicted attributes on MOT dataset before and after Stage2. It can be observed that Stage2 substantially improves the accuracy of predicted attributes. Therefore, fine-tuning with person ID labels refines the attributes on *U*. In the Stage3, the labeled *U* plus the original independent dataset will be combined for the final stage of fine-tuning. It thus can be inferred that, Stage2 helps the final stage of training by providing more accurate training data.

3.4. Fine-tuning on the combined dataset

The fine-tuning in previous stage produces more accurate attribute labels. We thus consider to combine the *T* and *U* for the final round of fine-tuning. As shown in Fig. 2, in the third stage, we first predict the attribute labels for dataset *U* with $\mathcal{O}^{S2}$. A new dataset labeled with attribute labels can hence be generated by merging *T* and *U*. Then, we fine-tune $\mathcal{O}^{S2}$ using the dataset *T&U* with a similar procedure in Stage 1. The fine-tuning outputs the final attribute detector $\mathcal{O}$.

For any test image, we can predict its *K*-dimensional attribute label with Eq. (1). In our implementation, we only select one attributes set as 1 in each type of attributes, and other attributes in this type set as 0. This essentially selects more accurate attributes and ensure the label incompatibility among attributes. Finally, $\mathcal{O}$ produces a sparse binary *K*-dimensional attribute vector. Our person ReID system uses this binary vector as feature and measures their distance with Cosine distance to identify the same person. The validity of this three-stage training procedure and the performance of selected attributes will be tested in Section 4.

4. Experiments

4.1. Datasets for training and testing

To conduct the first stage training, we choose the *PETA* [54] dataset as the training set. Each image in *PETA* is labeled with 61 binary attributes and 4 multi-class attributes. The 4 multi-class attributes are *footwear*, *hair*, *lowerbody* and *upperbody*, each of which has 11 color labels including *Black*, *Blue*, *Brown*, *Green*, *Grey*, *Orange*, *Pink*, *Purple*, *Red*, *White*, and *Yellow*, respectively. We hence expand 4 multi-class attributes into 44 binary attributes, resulting in a 105-dimensional binary attribute label.

To consider the incompatibility among attributes, we divide these 105 attributes into 15 types, including *Age*, *Gender*, *CarryObject*, *AccessoryObject*, *SleeveStyle*, *UpperStyle*, *UpperType*, *LowerStyle*, *LowerType*, *HairStyle*, *FootStyle*, *UpperColor*, *LowerColor*, *HairColor* and *FootColor*, respectively. More details can be found in Fig. 5. It should be noted that, we require each person to have at most one positive attribute within each type.

For the second stage training, we choose the *MOT challenge* [53] dataset to fine-tune dCNN $\mathcal{O}^{S1}$ with attributes triplet loss. *MOT challenge* is a dataset designed for multi-target tracking and provides the trajectories of each person. We thus could get the bounding box and ID label of each person. And we use more than 20,000 images on *MOT challenge*. Consequently, we will obtain more than 100,000 triplets.

To evaluate our model, we choose *VIPeR* [32], *PRID* [33], *GRID* [34], and *Market* [55] as test sets. Note that, *VIPeR*, *GRID* and *PRID* are included in the *PETA* dataset. When we test our algorithm on them, they will be excluded from the training set. For example, when we use the *VIPeR* for person ReID test, none of its images will be used for dCNN training.

CUHK03 is another popular dataset for Person ReID. However, *PETA* does not specify how many images in it are from *CUHK03*. Our statistical analysis shows that each testing group defined by *CUHK03* has about 20 IDs appear in *PETA*. Without the precise in-

Type	Attribute Label
Age	personalLess15, personalLess30, personalLess45, personalLess60, personalLarger60
Gender	personalFemale, personalMale
CarryObject	carryingBabyBuggy, carryingBackpack, carryingOther, carryingShoppingTro, carryingUmbrella, carryingFolder, carryingLuggageCase, carryingMessengerBag, carryingNothing, carryingPlasticBags, carryingSuitcase
AccessoryObject	accessoryHeadphone, accessoryHairBand, accessoryHat, accessoryKerchief, accessoryMuffler, accessoryNothing, accessorySunglasses
SleeveStyle	upperBodyNoSleeve, upperBodyShortSleeve, upperBodyLongSleeve
UpperStyle	upperBodyCasual, upperBodyFormal
UpperType	upperBodyJacket, upperBodyLogo, upperBodyPlaid, upperBodyThinStripes, upperBodySuit, upperBodySweater, upperBodyThickStripes, upperBodyTshirt, upperBodyOther, upperBodyVNeck
LowerStyle	lowerBodyCasual, lowerBodyFormal
LowerType	lowerBodyCapri, lowerBodyHotPants, lowerBodyJeans, lowerBodyLongSkirt, lowerBodyPlaid, lowerBodyThinStripes, lowerBodyShorts, lowerBodyShortSkirt, lowerBodySuits, lowerBodyTrousers
HairStyle	hairBald, hairShort, hairLong
FootStyle	footwearBoots, footwearLeatherShoes, footwearSandals, footwearShoes, footwearSneakers, footwearStocking
UpperColor	upperBodyBlack, upperBodyBlue, upperBodyBrown, upperBodyGreen, upperBodyGrey, upperBodyOrange, upperBodyPink, upperBodyPurple, upperBodyRed, upperBodyWhite, upperBodyYellow
LowerColor	lowerBodyBlack, lowerBodyBlue, lowerBodyBrown, lowerBodyGreen, lowerBodyGrey, lowerBodyOrange, lowerBodyPink, lowerBodyPurple, lowerBodyRed, lowerBodyWhite, lowerBodyYellow
HairColor	hairBlack, hairBlue, hairBrown, hairGreen, hairGrey, hairOrange, hairPink, hairPurple, hairRed, hairWhite, hairYellow
FootColor	footwearBlack, footwearBlue, footwearBrown, footwearGreen, footwearGrey, footwearOrange, footwearPink, footwearPurple, footwearRed, footwearWhite, footwearYellow

Fig. 5. Illustration of the 15 types of attributes.

formation provided by *PETA*, it is hard for us to manually remove all of the *CUHK03* images from *PETA*. Therefore, we test our methods on the above four test sets and do not use the *CUHK03* for testing.

4.2. Implementation details

We select the 16-layer VGG network [52] as our base dCNN architecture. We use the same kernel and filter sizes for all the hidden layers. We learn 105 binary attributes from *PETA*. When we fine-tune dCNN with attributes triplet loss, we follow the standard triplet loss algorithm [56] to select samples. Specifically, for each type, we first extract FC7 layer features for each image in U dataset. Then, we randomly select an image of a person as the *anchor* sample $u_{(a)}$. Another image of the same person with a large distance with $u_{(a)}$ is selected as *positive* sample $u_{(p)}$. An image of other persons with a smaller distance with $u_{(a)}$ is selected as *negative* sample $u_{(n)}$.

Parameters for learning are empirically set via cross-validation. The θ and γ in Eq. (4) are set as 1 and 0.01, respectively. We implement our approach with GTX TITAN X GPU, Intel i7 CPU, and 32GB memory. The first stage of training takes about one week, the second stage of fine-tuning takes about five days, and the third stage takes about three days.

4.3. Accuracy of predicted attributes and zero-shot learning

In the first experiment, we test the accuracy of predicted attributes on three datasets, *VIPeR, PRID* and *GRID*, as well as show the effects of combining different training stages. We select 1/10 of the whole training dataset for validation. For each attribute type, we show the top-1 and top-2 classification accuracies. Note that, when we test a certain dataset, images from this dataset wont appear in the training set. To test the convergence of our algorithm during the training stage, we also show the accuracies on the validation set. We summarize the results in Tables 1–3.

$Stage_1$ denotes the baseline dCNN $\mathcal{O}^{S1}$. $Stage_{1\&2}$ and $Stage^*_{1\&2}$ denote the updated dCNN $\mathcal{O}^{S2}$ after the second stage training using U and T. $Stage_{1\&3}$ first labels U with $\mathcal{O}^{S1}$, then combines U and T to fine-tune the $\mathcal{O}^{S1}$. WSMTAL denotes our final dCNN after the third stage training. From the experimental results, we can draw the following conclusions:

From Tables 1–3, we can draw the following conclusions:

1) Although $Stage_{1\&3}$ uses larger training set, it does not constantly outperform the baseline. This is because the expanded training data is labeled by $\mathcal{O}^{S1}$, and it does not provide new cues for fine-tuning $\mathcal{O}^{S1}$ in Stage 3.
2) $\mathcal{O}^{S2}$ produced by $Stage_{1\&2}$ does not constantly outperform baseline. This is reasonable because the goal of Stage 2 is to update the attribute labels of MOT dataset with the help of person ID labels, rather than updating the entire network and improving its discriminative power on unseen data, *e.g.*, testing data. This is why we only update the fully-connected layers in Stage 2 and keep the convolutional layers fixed. In another word, Stage 2 is important because it refines the attribute labels of MOT dataset U, thus the combined $U+T$ can be a better training set for Stage 3. Fig. 4 clearly shows that Stage 2 produces more accurate attribute labels.
3) WSMTAL is able to improve the accuracy of baseline by 3.0% in average on three datasets. This demonstrates our three-stage training framework can learn more robust semantic attributes. To intuitively show the accuracy of predicted attributes, we use the dCNN trained by WSMTAL to predict attributes on *MOT challenge* dataset. Some examples are illustrated in Fig. 6.
4) From the results on the validation set, it is obvious that $Stage_1$, $Stage_{1\&2}$, $Stage_{1\&3}$ and WSMTAL get comparable performance. This means that these algorithms converge well on the training set. It is also interesting to observe that WSMTAL performs not as good as $Stage_1$ and $Stage_{1\&3}$ on the validation set. This means further updating the deep model trained on T, i.e., $\mathcal{O}^{S1}$, with another dataset U drops the performance on validation set selected from T. However, with the help of additional person ID labels of U, WSMTAL is more suitable for Zero-shot learning and gets better generalization ability. WSMTAL also achieves the best performance on *VIPeR, PRID*, and *GRID*.
5) To show that extra person ID labels, i.e., the MOT challenge dataset, help our model training, we compare the performance

Table 1
The classification accuracies of attributes of the first 5 types on the *VIPeR*, *GRID* and *PRID* datasets.

Types	Number		Validation(%)		VIPeR(%)		PRID(%)		GRID(%)	
			Top-1	Top-2	Top-1	Top-2	Top-1	Top-2	Top-1	Top-2
Age	5	$Stage_1$	86.41	97.03	44.34	80.42	48.13	77.35	42.56	78.35
		$Stage_{1\&2}$	84.56	95.18	41.09	77.85	46.03	75.45	39.87	76.42
		$Stage^*_{1\&2}$	**87.16**	96.18	44.09	81.85	48.15	77.61	43.22	78.42
		$Stage_{1\&3}$	86.47	**97.51**	45.02	80.33	48.23	78.24	42.45	77.99
		WSMTAL	86.02	96.87	**46.99**	**81.84**	**50.43**	**79.13**	**45.13**	**81.84**
Sex	2	$Stage_1$	96.53	**100.00**	70.05	**100.00**	67.50	**100.00**	67.82	**100.00**
		$Stage_{1\&2}$	95.47	**100.00**	69.12	**100.00**	66.43	**100.00**	66.34	**100.00**
		$Stage^*_{1\&2}$	**97.01**	**100.00**	70.03	**100.00**	66.32	**100.00**	68.43	**100.00**
		$Stage_{1\&3}$	96.49	**100.00**	70.13	**100.00**	67.66	**100.00**	68.03	**100.00**
		WSMTAL	95.83	**100.00**	**71.20**	**100.00**	**69.64**	**100.00**	**69.73**	**100.00**
CarryObject	11	$Stage_1$	**86.35**	96.31	26.46	46.82	29.46	52.83	26.96	41.28
		$Stage_{1\&2}$	84.75	95.92	25.12	47.03	28.22	50.37	25.47	41.96
		$Stage^*_{1\&2}$	86.27	**96.55**	25.88	45.96	28.39	51.72	26.05	40.78
		$Stage_{1\&3}$	86.32	95.69	27.19	46.71	30.11	53.07	26.88	41.63
		WSMTAL	85.08	94.24	**29.27**	**49.01**	**31.71**	**54.82**	**28.74**	**44.53**
AccessoryObject	7	$Stage_1$	91.86	97.54	44.32	66.33	57.66	75.58	52.45	83.77
		$Stage_{1\&2}$	90.25	96.44	43.21	65.74	56.41	74.93	51.68	83.40
		$Stage^*_{1\&2}$	**92.18**	**98.54**	43.52	66.08	55.57	75.33	52.63	84.29
		$Stage_{1\&3}$	91.55	97.19	44.83	66.85	57.38	75.69	51.85	83.79
		WSMTAL	90.62	97.40	**46.17**	**69.10**	**60.14**	**78.85**	**54.19**	**86.06**
SleeveStyle	3	$Stage_1$	99.14	**100.00**	80.17	93.73	85.37	95.62	42.71	77.96
		$Stage_{1\&2}$	98.75	**100.00**	79.64	93.46	84.33	94.23	42.22	77.68
		$Stage^*_{1\&2}$	99.42	**100.00**	79.98	92.96	85.67	95.39	43.01	78.25
		$Stage_{1\&3}$	**99.93**	**100.00**	79.38	93.63	84.96	95.73	43.09	76.93
		WSMTAL	98.96	**100.00**	**85.76**	**98.73**	**87.09**	**97.88**	**45.13**	**81.84**

Table 2
The classification accuracies of attributes of the next 7 types on the *VIPeR*, *GRID* and *PRID* datasets.

Types	Number		Validation(%)		VIPeR(%)		PRID(%)		GRID(%)	
			Top-1	Top-2	Top-1	Top-2	Top-1	Top-2	Top-1	Top-2
UpperStyle	2	$Stage_1$	99.01	**100.00**	90.17	**100.00**	82.64	**100.00**	80.53	**100.00**
		$Stage_{1\&2}$	97.85	**100.00**	89.25	**100.00**	82.01	**100.00**	80.11	**100.00**
		$Stage^*_{1\&2}$	**99.53**	**100.00**	89.66	**100.00**	82.23	**100.00**	79.85	**100.00**
		$Stage_{1\&3}$	98.69	**100.00**	90.33	**100.00**	82.49	**100.00**	81.36	**100.00**
		WSMTAL	98.96	**100.00**	**92.96**	**100.00**	**84.10**	**100.00**	**83.63**	**100.00**
UpperType	10	$Stage_1$	85.31	99.36	62.57	83.71	57.48	79.13	55.97	76.53
		$Stage_{1\&2}$	84.27	99.17	61.48	83.42	57.39	79.11	55.48	75.68
		$Stage^*_{1\&2}$	**86.51**	99.97	62.37	83.59	58.95	81.24	53.82	73.35
		$Stage_{1\&3}$	86.33	**100.00**	61.34	83.66	57.24	78.54	54.76	76.42
		WSMTAL	84.90	98.44	**64.34**	**85.93**	**60.81**	**81.56**	**58.99**	**79.83**
LowerStyle	2	$Stage_1$	96.27	**100.00**	90.23	**100.00**	81.97	**100.00**	79.19	**100.00**
		$Stage_{1\&2}$	95.69	**100.00**	89.96	**100.00**	81.05	**100.00**	78.56	**100.00**
		$Stage^*_{1\&2}$	96.39	**100.00**	88.45	**100.00**	79.32	**100.00**	77.73	**100.00**
		$Stage_{1\&3}$	**96.42**	**100.00**	90.15	**100.00**	82.01	**100.00**	79.08	**100.00**
		WSMTAL	96.33	**100.00**	**93.24**	**100.00**	**84.44**	**100.00**	**81.70**	**100.00**
LowerType	10	$Stage_1$	93.55	**100.00**	70.39	90.14	64.32	89.44	41.75	64.35
		$Stage_{1\&2}$	92.29	99.42	70.11	90.27	63.57	89.28	40.19	63.70
		$Stage^*_{1\&2}$	**95.94**	**100.00**	68.32	87.16	62.70	86.48	40.95	61.43
		$Stage_{1\&3}$	93.27	99.33	69.74	89.36	64.68	88.93	41.38	63.77
		WSMTAL	91.67	98.96	**74.92**	**95.94**	**66.20**	**91.66**	**44.04**	**67.03**
HairStyle	3	$Stage_1$	96.79	**100.00**	68.59	94.19	70.13	94.53	66.23	95.43
		$Stage_{1\&2}$	96.23	**100.00**	67.13	94.00	68.47	94.53	66.12	94.13
		$Stage^*_{1\&2}$	96.88	**100.00**	67.36	93.15	69.57	93.84	64.33	93.07
		$Stage_{1\&3}$	**96.93**	**100.00**	68.48	94.32	69.88	94.89	67.05	95.00
		WSMTAL	96.88	**100.00**	**71.42**	**95.91**	**71.51**	**96.81**	**68.82**	**96.03**

between using and without using MOT challenge dataset, respectively. In Tables 1–3, $Stage^*_{1\&2}$ denotes the performance of dCNN trained without using MOT dataset. It can be seen that, our WSMTAL outperforms $Stage^*_{1\&2}$ in most cases. It is also interesting to notice that, $Stage^*_{1\&2}$ constantly outperforms $Stage_{1\&2}$ on the dataset *T*, where the validation is from. This is reasonable because the model of $Stage_{1\&2}$ is optimized on another domain *U*, i.e., MOT dataset, and targets to refine the attribute labels of MOT dataset, rather than to improve the discriminative power on the dataset *T*. Differently, $Stage^*_{1\&2}$ is directly optimized on the dataset *T*, thus shows better performance than $Stage_{1\&2}$.

4.4. Performance on two-camera datasets

This experiment tests deep attributes on two-camera person ReID tasks. Three datasets are employed. 10 random tests are first performed for each dataset. Then, the average *Cumulative Match*

Table 3
The classification accuracies of attributes of the last 5 types on *VIPeR*, *GRID* and *PRID* datasets.

Types	Number		Validation(%)		VIPeR(%)		PRID(%)		GRID(%)	
			Top-1	Top-2	Top-1	Top-2	Top-1	Top-2	Top-1	Top-2
FootStyle	6	$Stage_1$	71.53	95.06	35.38	64.63	45.34	72.53	30.15	55.91
		$Stage_{1\&2}$	71.59	94.69	33.62	63.92	44.43	72.04	29.96	56.82
		$Stage^*_{1\&2}$	**72.86**	**97.15**	33.78	62.14	43.84	70.08	27.29	53.11
		$Stage_{1\&3}$	71.66	95.88	34.96	64.55	46.02	73.10	30.05	56.39
		WSMTAL	70.83	94.79	**38.36**	**66.25**	**47.75**	**74.14**	**33.80**	**60.13**
UpperColor	11	$Stage_1$	79.35	**93.23**	44.33	67.11	28.21	51.03	30.25	52.47
		$Stage_{1\&2}$	78.02	91.68	42.96	66.49	27.63	50.41	30.23	51.39
		$Stage^*_{1\&2}$	78.89	92.96	44.65	67.99	26.38	48.17	29.65	51.03
		$Stage_{1\&3}$	**79.44**	91.99	45.36	66.52	28.04	50.88	31.18	53.00
		WSMTAL	78.12	91.15	**47.92**	**69.13**	**28.97**	**52.31**	**32.90**	**55.04**
LowerColor	11	$Stage_1$	95.62	98.96	47.76	72.36	38.06	70.53	50.91	76.35
		$Stage_{1\&2}$	94.18	98.78	46.33	71.45	37.82	69.65	50.13	76.42
		$Stage^*_{1\&2}$	**96.77**	99.12	46.55	70.49	36.24	70.12	47.69	73.83
		$Stage_{1\&3}$	94.37	**100.00**	48.25	71.88	38.27	70.48	49.97	75.99
		WSMTAL	94.23	98.44	**51.75**	**75.69**	**40.99**	**72.82**	**52.18**	**77.89**
HairColor	11	$Stage_1$	96.39	99.13	55.37	76.14	48.11	80.37	44.35	64.50
		$Stage_{1\&2}$	95.44	98.27	54.23	75.38	46.27	81.04	43.19	65.39
		$Stage^*_{1\&2}$	96.11	**99.73**	52.06	71.89	47.56	79.63	41.83	63.20
		$Stage_{1\&3}$	**96.50**	98.69	56.30	76.23	48.09	82.10	43.96	66.42
		WSMTAL	95.84	97.92	**57.74**	**79.30**	**50.80**	**84.49**	**47.09**	**70.71**
FootColor	11	$Stage_1$	93.68	98.66	44.28	68.13	62.45	83.69	30.30	60.13
		$Stage_{1\&2}$	92.71	96.75	43.19	67.48	60.97	82.47	28.56	59.48
		$Stage^*_{1\&2}$	**94.80**	**99.17**	42.24	66.90	61.01	80.73	29.05	58.14
		$Stage_{1\&3}$	93.07	97.58	45.07	68.25	62.37	83.02	30.66	59.88
		WSMTAL	92.19	96.88	**46.03**	**70.84**	**67.14**	**84.81**	**34.48**	**63.80**

Characteristic (CMC) curves of these tests are calculated and used for performance evaluation. The experimental settings on three datasets are introduced as follows:

VIPeR : 632 persons are included in the *VIPeR* dataset. Two images with size 48×128 of each person are taken by camera A and camera B, respectively in different scenarios of illumination, postures and viewpoints. Different from most of existing algorithms, our WSMTAL does not need training on the target dataset. To make fair comparison with other algorithms, we use similar settings for performance evaluation, i.e., randomly selecting 10 test sets, and each contains 316 persons.

PRID : This dataset is specially designed for person ReID in single shot. It contains two image sets containing 385 and 749 persons captured by camera A and camera B, respectively. These two datasets share 200 persons in common. For the purpose of fair comparison with other algorithms, we follow the protocol in [33], and create a probe set and a gallery set, where all training samples are excluded. The probe set includes images of 100 persons from camera A. The gallery set is made up of images from 649 persons capture by camera B.

GRID : This dataset includes images collected by 8 non-adjacent cameras fixed at a subway station. The probe set contains images of about 250 persons. The gallery set contains images of about 1025 persons, among which 775 persons do not match anyone in the probe set. For the purpose of fair comparison, images of 125 persons shared by the two sets are employed for training. The remaining 125 persons and 775 distracters are used for the testing.

Compared Algorithms : We compare our approach with many recent works. Compared works that learn distance metrics for person ReID include RPML [10], PRDC [17], RSVM [65], Salmatch [57], LMF [58], PCCA [9], KISSME [13], kLFDA [14], KCCA [59], TSR [60], EPKFM [19],LOMO + XQDA [20],MRank-PRDC [34], MRank-RSVM [34], RQDA [66], MLAPG [23], CSL [22] and LDNS [61]. Compared works based on traditional attribute learning are AIR [26], OAR [28] ,LOREA [31] and JLSAT [62]. Related works that leverage deep learning include DML [50], IDLA [51], Deep-RDC [24], DGDropout [64], Gate S-CNN [63] and Deep-TCP [38]. The compared CMC scores at different ranks on three datasets are shown in Tables 4, 5, and 6, respectively.

The three tables clearly show that, even it is not fine-tuned with extra data, the baseline dCNN $\mathcal{O}^{S1}$ achieves fairly good results on three datasets, especially on *PRID* and *GRID*. Additionally, if we fine-tune the baseline dCNN using our attributes triplet loss, we achieve an additional 3.5% improvement at rank 1 on *VIPeR*, 2.8% on *PRID*, and 1.7% on *GRID*, respectively. This indicates that our three-stage training framework improves the performance by progressively adding more information into the training procedure.

Our WSMTAL algorithm has surpassed many existing algorithms on the *VIPeR*, *PRID* and *GRID* datasets. Some recent works like AIR [26], OAR [28], LOREA [31], and JLSAT [62] also learn attributes for person ReID. The comparison in Table 4 clearly shows the advantages of our deep model in attribute prediction. Some previous works like DML [50], IDLA [51], Deep-RDC [24], DG-Dropout [64], Gate S-CNN [63] and Deep-TCP [38] take advantages of deep learning in person ReID. Different from them, our work generates camera-independent mid-level attributes, which can be used as discriminative features for identifying persons on different datasets. The experiments results in Table 4 also show that our method outperforms these works.

Because we use the predicted binary attributes as features for person ReID, we can also learn a distance metric to further improve the ReID accuracy. We select XQDA [20] for the distance metric learning. As can be seen from three tables, our approach with XQDA [20], i.e., WSMTAL + XQDA, achieves better performance than WSMTAL. This clearly shows that our work can easily combine with existing distance metric learning works to further boost the performance.

4.5. Performance on multi-camera dataset

We further test our approach in a more challenging multi-camera person ReID task. We employ the *Market* dataset [55], where more than 25,000 images of 1501 labeled persons are collected from 6 cameras. Each person has 17 images in average,

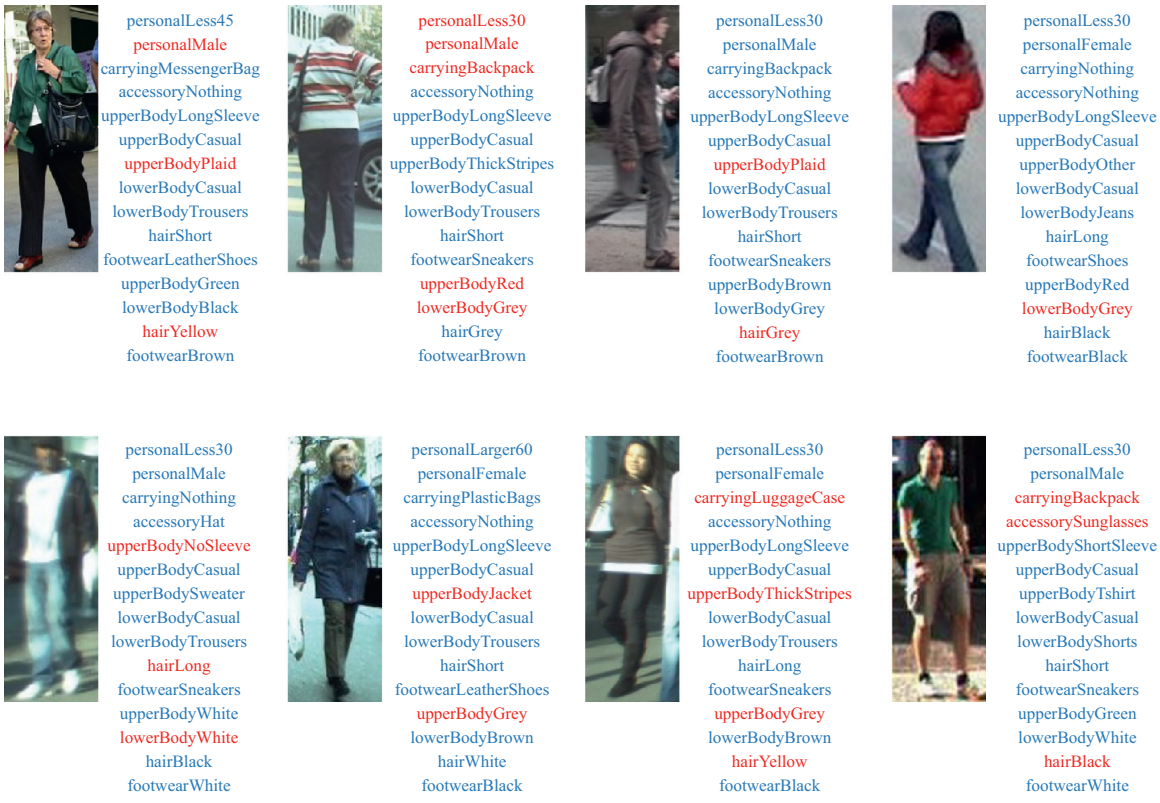

Fig. 6. Examples of predicted attributes on *MOT challenge* by the learned dCNN after three stages of training. Texts with blue color are correct attributes, while those with red color are false attributes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
CMC scores, i.e., percentage (%) of correct matches, of ranks 1, rank 5, rank 10, rank 20 on the *VIPeR* dataset.

Methods		Rank 1	Rank 5	Rank 10	Rank 20
Metric Learning based ReID	RPML [10]	27.0	57.0	69.0	83.0
	Salmatch [57]	30.2	52.4	65.5	79.1
	LMF [58]	29.1	52.3	65.9	80.0
	KISSME [13]	19.6	47.5	62.2	77.0
	KCCA [59]	37.3	71.4	84.6	92.3
	kLFDA [14]	32.2	65.8	79.7	90.9
	LOMO + XQDA [20]	40.0	68.9	81.5	91.1
	CSL [22]	34.8	68.7	82.3	91.8
	MLAPG [23]	40.7	69.9	82.3	92.4
	TSR [60]	31.6	68.6	82.8	**94.6**
	EPKFM [19]	36.8	70.4	83.7	91.7
	LDNS [61]	42.3	71.4	82.9	92.1
Attributes Learning based ReID	AIR [26]	18.0	38.8	51.1	71.2
	LLCNN-P [47,48]	13.89	34.02	47.41	-
	OAR [28]	21.4	41.5	55.2	71.5
	LORAE [31]	42.3	72.2	81.6	89.6
	JLSAT [62]	45.4	–	–	–
Deep Learning based ReID	IDLA [51]	34.8	54.3	76.5	87.6
	DML [50]	28.2	59.3	73.5	86.4
	Deep-RDC [24]	40.5	60.8	70.4	84.4
	Gate S-CNN [63]	37.8	66.9	76.3	–
	DGDropout [64]	38.6	–	–	–
	Deep-TCP [38]	**47.8**	**74.7**	**84.8**	91.1
Proposed	SSDAL [41]	37.9	65.5	75.6	88.4
	$Stage_1$	36.2	63.9	73.5	84.7
	$Stage_{1\&2}$	36.3	63.1	72.9	82.4
	$Stage_{1\&3}$	37.3	61.5	72.7	81.5
	WSMTAL	39.7	66.9	76.5	86.6
	WSMTAL + XQDA	47.1	71.5	80.3	88.2

Table 5
CMC scores, i.e., percentage (%) of correct matches, of ranks 1, rank5, rank 10, rank 20 on the *PRID* dataset.

Methods	Rank 1	Rank 5	Rank 10	Rank 20
RPML [10]	4.8	14.3	21.6	30.2
PRDC [17]	4.5	12.6	19.7	29.5
RSVM [65]	6.8	16.5	22.7	31.5
Salmatch [57]	4.9	17.5	26.1	33.9
LMF [58]	12.5	23.9	30.7	36.5
PCCA [9]	3.5	10.9	17.9	27.1
KISSME [13]	4.1	12.8	21.1	31.8
kLFDA [14]	7.6	18.9	25.6	37.4
KCCA [59]	14.5	34.3	46.7	59.1
LOREA [31]	18.0	37.4	50.1	66.6
LOMO + XQDA [20]	15.3	35.7	41.2	53.8
MLAPG [23]	16.6	33.1	41.4	52.5
JLSAT [62]	26.8	–	–	–
Deep-TCP [38]	22.0	–	47.0	57.0
LDNS [61]	**29.80**	**52.9**	**66.0**	**76.5**
SSDAL [41]	20.1	47.4	55.7	68.6
$Stage_1$	19.6	46.7	55.1	66.4
$Stage_{1\&2}$	19.4	46.4	53.5	66.2
$Stage_{1\&3}$	20.8	43.0	55.1	67.3
WSMTAL	22.4	47.8	56.8	67.6
WSMTAL + XQDA	24.4	52.3	62.5	74.2

Table 6
CMC scores, i.e., percentage (%) of correct matches, of ranks 1, rank5, rank 10, rank 20 on the *GRID* dataset.

Methods	Rank 1	Rank 5	Rank 10	Rank 20
PRDC [17]	9.7	22.0	33.0	44.3
RSVM [65]	10.2	24.6	33.3	43.7
MRank-PRDC [34]	11.1	26.1	35.8	46.6
MRank-RSVM [34]	12.2	27.8	36.3	49.3
RQDA [66]	15.2	30.1	39.2	49.3
EPKFM [19]	16.3	35.8	46.0	57.6
LOMO + XQDA [20]	16.6	35.4	41.8	52.4
LLCNN-P [47,48]	18.32	**46.16**	**62.56**	–
SSDAL [41]	19.1	35.6	48.0	58.4
$Stage_1$	17.5	34.5	42.8	55.3
$Stage_{1\&2}$	16.8	32.0	43.3	57.5
$Stage_{1\&3}$	17.5	35.4	44.5	55.8
WSMTAL	19.2	38.1	47.8	58.7
WSMTAL + XQDA	**23.4**	39.6	49.8	**60.3**

Table 7
CMC scores of ranks 1 and mean Average Precision (mAP) on the *Market* dataset. Numbers indicate the percentage (%) of correct matches within a specific rank.

Single Query	Rank 1	mAP
Salmatch [57]	20.5	8.2
SDALF [1]	33.5	13.5
BGG [55]	34.4	14.1
KISSME [13]	40.5	19.0
MFA [14]	45.7	18.2
kLFDA [14]	51.3	24.4
LOMO + XQDA [20]	43.8	22.2
LDNS [61]	55.4	29.9
Gate S-CNN [63]	**65.9**	**39.6**
SSDAL [41]	39.4	19.6
WSMTAL	49.5	29.2
Multiple Query	Rank 1	MAP
BGG+MultiQ_max [55]	42.1	18.5
kLFDA [14]	52.7	27.4
LOMO + XQDA [20]	54.1	28.4
LDNS [61]	68.0	41.9
Gate S-CNN [63]	**76.0**	**48.5**
SSDAL [41]	49.0	25.8
WSMTAL	56.6	31.2

which show substantially different appearances due to variances of viewpoints, illumination, backgrounds, *etc*. This dataset is also larger than most of existing person ReID datasets. Because *Market* has clearly provided the training set, we use images in the training set and their person ID labels to fine-tune our dCNN $\mathcal{O}^{S2}$.

In contrast to the two-camera person ReID task, the multi-camera person ReID targets to identify the query person across image sets from multiple cameras. Therefore, our task is to query and rank all images from these cameras, according to the given probe image (i.e., Single Query) or tracklet (i.e., Multiple Query) of a person. Because this process is similar to image retrieval, we evaluate the performance by mean Average Precision (mAP) and accuracy at Rank 1, following the protocol in [55]. The results are shown in Table. 7. More details about feature pooling can be found in [55].

From Table 7, we can observe that our approach outperforms most of the compared methods for both single query and multi-query scenarios in mAP. Our method does not perform as good as the latest LDNS [61] and Gate S-CNN [63] methods on *Market* dataset. Note that, both of LDNS [61] and Gate S-CNN [63] train their models directly on the training set of the Market dataset. Therefore, the underlying reason maybe because our method simply transfers the learned low-dimensional attribute features from an independent and relative small dataset *PETA* to the large Market dataset. Moreover, the bounding box annotations in the Market dataset are generated from the DPM [67] detection model, which differ from the manually annotated bounding boxes in *PETA*. On datasets like *VIPER*, where the bounding boxes are also manually annotated, our method performs better than LDNS [61] and Gate S-CNN [63] in Table 4.

4.6. Discussions

In this part, we further discuss some interesting aspects of our method that may have been missed in the above experimental evaluations.

It should be noted that, our training and testing sets are from different domains. To be specific, the attribute prediction model is trained on PETA and MOT Challenge, rather than the training sets defined by VIPeR, PRID, and GRID. This setting is thus more challenging than the commonly used one in compared works, i.e., the training set and testing set are both from the same domain or dataset. Under such setting, our method still shows competitive performance in Tables 4–7. Our work also shows reasonable performance when is compared with recent works. After using XQDA, our rank-1 accuracy on GRID is only 0.8% lower than the accuracy reported in [68]. Our algorithm also outperforms the method of Matsukawa *et al.* using pixel feature [69] on both *VIPeR* and *GRID*, i.e., 22.8% [69] vs. our 23.4% on GRID, and 42.3 [69] % vs. our 47.1% on VIPeR. Those experimental results show our method has a strong feature generalization ability, i.e., attribute feature is trained on one set but gets reasonable performance on other independent testing sets. This could be valuable for real applications, where the attribute training sets on target domain could be hard to collect.

By using attributes features of only 105 dimensions, our method achieves promising performance on four public datasets. It is interesting to see the ReID performance after combining the compact attribute features and classic visual features. To verify this point, we integrate the appearance-based features with attributes features for better discriminative power. Table 8 shows the performance of fusing deep attributes with appearance-based feature LOMO [20], i.e., LOMO + XQDA + WSMTAL. It is obvious that fusing appearance-based features further improves WSMTAL, *e.g.*, CMC score achieves 45.3 at Rank-1. Therefore, combining with visual feature would further ensure the performance of attributes features in real applications.

Table 8
Additional experimental results on *VIPeR*.

Method	Rank 1	Rank 5	Rank 10	Rank 20
WSMTAL	39.7	66.9	76.5	86.6
WSMTAL + XQDA	47.1	71.5	80.3	88.2
LOMO + WSMTAL + XQDA	51.3	78.2	85.1	90.2
FC7 fine-tuned on T	26.5	48.2	61.1	72.3
FC7 fine-tuned on U	10.1	21.6	31.7	45.3
FC7 fine-tuned on $T+U$	27.4	49.7	62.3	74.4

Many image retrieval works use the output of FC-7 layer in VGG-16 as image feature. Therefore, another way of learning mid-level feature for person ReID is fine-tunning the FC7 layer with triplet loss similar to the one in WSMTAL, i.e., updating the dCNN to make same person have similar FC-7 layer features and vice versa. The FC7 features learned in this way are also not limited to the 105 dimensions, thus might be more discriminative than attributes. To test the validity of this strategy, we fine-tune the FC7 layer of VGG-Net using person ID labels on different datasets, i.e., *T, U*, and $T+U$, respectively. Experimental results in Table 8 clearly indicates that that deep attributes outperforms such FC7 features. This clearly validates the contribution and importance of attributes.

Compared with our conference paper, which uses AlexNet, WSMTAL uses deeper VGG-Net. The baseline performance of VGG-Net is about 2% higher than the one of AlexNet. Table 8 shows that our WSMTAL framework is about 12% better than those produced by directly using VGG-Net for feature learning in Rank 1 on VIPeR. This also shows that our proposed weakly supervised method brings more significant performance gain than the use of a deeper network.

There are some works that use deep learning to recognize pedestrian attributes [44,46–48]. The DeepMAR [44] is a deep attributes learning model which can learn the attributes correlations. And the work by Yu et al. [46] proposes a weakly supervised deep learning model to recognize attributes and infer the locations of attributes. Meanwhile, a multi-label convolutional neural network (MLCNN) [47,48] is formulated to predict multiple attributes with body part division. Although most of those works are not working on person ReID, they can be important references for our work. For example, those works show that considering body parts, locations, correlations of attributes may further improve the attribute prediction accuracy. Referring to those works, we will add more helpful information to our weakly supervised attribute learning model. This will be investigated in our future work.

5. Conclusions and future work

This paper addresses the person ReID problem using deeply learned human attribute features. We propose a novel Weakly supervised Multi-Type Attribute Learning (WSMTAL) algorithm, which considers the contextual cues among attributes and progressively boosts the accuracy of attributes only using a limited number of labeled data. Our attributes triplet loss makes it possible to use images only with person ID labels for training attribute detectors in a dCNN framework. Extensive experiments on four benchmark datasets demonstrate that our method performs reasonably good in attribute detection and outperforms many recent person ReID methods. Moreover, our algorithm needs no further training on the target datasets. It means that once the attribute prediction dCNN model is trained, it can be applied in person ReID tasks on different datasets. The dCNN model fine-tuning only requires images with person ID labels, which can be easily obtained by Multi-target Tracking algorithms. Further considering the spatial locations of attributes might improve the accuracy of attribute detection. These would be our future work.

Acknowledgments

This work is supported by National Science Foundation of China under Grant No. 61572050, 61672519, 91538111, 61620106009, 61429201, and the National 1000 Youth Talents Plan, in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

References

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, CVPR, 2010.
[2] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, BMVC, 2011.
[3] B. Ma, Y. Su, F. Jurie, Bicov: a novel image representation for person re-identification and face verification, BMVC, 2012.
[4] C. Liu, S. Gong, C.C. Loy, X. Lin, Person re-identification: what features are important? ECCV, 2012.
[5] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, CVPR, 2013.
[6] X. Wang, R. Zhao, Person re-identification: System design and evaluation overview, in: Person Re-Identification, 2014, pp. 351–370.
[7] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, Q. Tian, Query-adaptive late fusion for image search and person re-identification, CVPR, 2015.
[8] A.J. Ma, P.C. Yuen, J. Li, Domain transfer support vector ranking for person re-identification without target camera label information, ICCV, 2013.
[9] M. Dikmen, E. Akbas, T.S. Huang, N. Ahuja, Pedestrian recognition with a learned metric, ACCV, 2011.
[10] M. Hirzer, P.M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, ECCV, 2012.
[11] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, CVPR, 2013.
[12] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, in: PAMI, volume 29, 2007, pp. 40–51.
[13] M. Köstinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: CVPR.
[14] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, ECCV, 2014.
[15] C. Liu, C.C. Loy, S. Gong, G. Wang, Pop: person re-identification post-rank optimisation, ICCV, 2013.
[16] Z. Li, S. Chang, F. Liang, T.S. Huang, L. Cao, J.R. Smith, Learning locally-adaptive decision functions for person verification, CVPR, 2013.
[17] W.-S. Zheng, S. Gong, T. Xiang, Re-identification by relative distance comparison, CVPR, 2013.
[18] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, ECCV, 2014.
[19] D. Chen, Z. Yuan, G. Hua, N. Zheng, J. Wang, Similarity learning on an explicit polynomial kernel feature map for person re-identification, CVPR, 2015.
[20] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, CVPR, 2015.
[21] Y.-C. Chen, W.-S. Zheng, J. Lai, Mirror representation for modeling view-specific transform in person re-identification, IJCAI, 2015.
[22] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, J. Wang, Person re-identification with correspondence structure learning, ICCV, 2015.
[23] S. Liao, S.Z. Li, Efficient psd constrained asymmetric metric learning for person re-identification, ICCV, 2015.
[24] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognit. 48 (10) (2015) 2993–3003.
[25] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, Y. Tian, Unsupervised cross-dataset transfer learning for person re-identification, CVPR, 2016.
[26] R. Layne, T.M. Hospedales, S. Gong, Q. Mary, Person re-identification by attributes., BMVC, 2012.
[27] R. Layne, T.M. Hospedales, S. Gong, Towards person identification and re-identification with attributes, ECCV Workshops, 2012.
[28] R. Layne, T.M. Hospedales, S. Gong, Attributes-based re-identification, in: Person Re-Identification, Springer, 2014, pp. 93–117.
[29] R. Layne, T.M. Hospedales, S. Gong, Re-id: hunting attributes in the wild, BMVC, 2014.
[30] C. Su, F. Yang, G. Zhang, Q. Tian, W. gao, L. Davis, Tracklet-to-tracklet person re-identification by attributes with discriminative latent space mapping, ICMS, 2015.
[31] C. Su, F. Yang, S. Zhang, Q. Tian, L.S. Davis, W. Gao, Multi-task learning with low rank attribute embedding for person re-identification, ICCV, 2015.
[32] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, PETS, 2007.
[33] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Image Analysis, Springer, 2011, pp. 91–102.
[34] C.C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, 2013.
[35] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, NIPS, 2012.

[36] R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation, PAMI, 2015.
[37] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, CVPR, 2014.
[38] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, CVPR, 2016.
[39] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification.
[40] E. Ustinova, Y. Ganin, V. Lempitsky, Multiregion bilinear convolutional neural networks for person re-identification, 2015.
[41] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, ECCV, 2016.
[42] S. Shankar, V.K. Garg, R. Cipolla, Deep-carving: Discovering visual attributes by carving deep neural nets, CVPR, 2015.
[43] Q. Chen, J. Huang, R. Feris, L.M. Brown, J. Dong, S. Yan, Deep domain adaptation for describing people based on fine-grained clothing attributes, CVPR, 2015.
[44] D. Li, X. Chen, K. Huang, Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios, in: Pattern Recognition (ACPR),year 2015 3rd IAPR Asian Conference on, IEEE, 2015, pp. 111–115.
[45] J. Huang, R.S. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, ICCV, 2015.
[46] K. Yu, B. Leng, Z. Zhang, D. Li, K. Huang, Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization, ArXiv preprint. arXiv:1611.05603.
[47] J. Zhu, S. Liao, D. Yi, Z. Lei, S.Z. Li, Multi-label cnn based pedestrian attribute learning for soft biometrics, in: International Conference on Biometrics, 2015, pp. 535–540.
[48] J. Zhu, S. Liao, Z. Lei, S.Z. Li, Multi-label convolutional neural network based pedestrian attribute classification, Image and Vision Computing, 2016.
[49] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, CVPR, 2014.
[50] D. Yi, Z. Lei, S.Z. Li, Deep metric learning for practical person re-identification, ICPR, 2014.
[51] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, CVPR, 2015.
[52] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014.
[53] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, K. Schindler, Towards a benchmark for multi-target tracking, in: arXiv:1504.01942 2015. arXiv preprint, Motchallenge year 2015.
[54] Y. Deng, P. Luo, C.C. Loy, X. Tang, Pedestrian attribute recognition at far distance, ACM MM, 2014.
[55] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, ICCV, 2015.
[56] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, CVPR, 2015.
[57] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, ICCV, 2013.
[58] R. Zhao, W. Ouyang, X. Wang, Learning midlevel filters for person reidentification, CVPR, 2014.
[59] G. Lisanti, I. Masi, A.D. Bimbo, Matching people across camera views using kernel canonical correlation analysis, ICDSC, 2014.
[60] Z. Shi, T.M. Hospedales, T. Xiang, Transferring a semantic representation for person re-identification and search, CVPR, 2015.
[61] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, CVPR, 2016.
[62] P. Peng, Y. Tian, T. Xiang, Y. Wang, T. Huang, Joint learning of semantic and latent attributes, ECCV, 2016.
[63] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, ECCV, 2016.
[64] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, CVPR, 2016.
[65] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking, BMVC, 2010.
[66] S. Liao, Y. Hu, S.Z. Li, Joint dimension reduction and metric learning for person re-identification, in: arXiv preprint, arXiv:1406.4216 2014.
[67] P.F. Felzenszwalb, R.B. Girshick, D.M. Allester, D. Ramanan, Object detection with discriminatively trained part based models 32 (2010) 1627–1645.
[68] D. Chen, Z. Yuan, Z. Yuan, B. Chen, N. Zheng, Similarity learning with spatial constraints for person re-identification, CVPR, 2016.
[69] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical gaussian descriptor for person re-identification, CVPR, 2016.

Chi Su is working toward the PhD degree in the Institute of Digital Media, EECS, Peking University. His research includes computer vision and machine learning, with focus on object detection, object tracking, and human identification and recognition. He is a student member of the IEEE.

Shiliang Zhang received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2012. He was a Post-Doctoral Scientist with NEC Laboratories America and a Post-Doctoral Research Fellow with The University of Texas at San Antonio. He is currently a tenure-track Assistant Professor with the School of Electronic Engineering and Computer Science, Peking University. He has authored or co-authored over 40 papers in journals and conferences, including IEEE T-PAMI, T-IP, T-MM, T-CSVT, Pattern Recognition, ACM Multimedia, and ICCV. His research interests include large-scale image retrieval and computer vision for autonomous driving.
He was a recipient of the National 1000 Youth Talents Plan of China, the Outstanding Doctoral Dissertation Awards from the Chinese Academy of Sciences and Chinese Computer Federation, the President Scholarship from the Chinese Academy of Sciences, the NEC Laboratories America Spot Recognition Award, and the Microsoft Research Fellowship. He was a recipient of the Top 10

Junliang Xing received his dual B.S. degrees in computer science and mathematics from Xi'an Jiaotong University, Shaanxi, China, in 2007, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2012. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His currently research interests mainly focus on solving computer vision and machine learning problems using deep neural network models.

Wen Gao received the PhD degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He was a professor of computer science in the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a professor in the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is currently a professor of computer science with Peking University, Beijing. He has authored extensively, including five books and more than 600 technical articles in refereed journals and conference proceedings in image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics. He served or serves on the editorial board for several journals, such as the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Multimedia, the IEEE Transactions on Image Processing, the IEEE Transactions on Autonomous Mental Development, the EURASIP Journal of Image Communications, and the Journal of Visual Communication and Image Representation. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE International Conference on Multimedia & Expo and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations. He is a fellow of the IEEE and ACM.

Qi Tian received the PhD degree in ECE from the University of Illinois at Urbana-Champaign (UIUC), in 2002. He is currently a professor in the Department of Computer Science, University of Texas at San Antonio (UTSA). He was a tenured associate professor from 2008 to 2012 and a tenure-track assistant professor from 2002 to 2008.
His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics and published more than 350 refereed journal and conference papers. He received the Best Paper Awards in ACM ICMR 2015, PCM 2013, MMM 2013, and ACM ICIMCS 2012, a Top 10 percent Paper Award in MMSP 2011, a Student Contest Paper Award in ICASSP 2006. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, Blippar, SALSI, Akiira Media Systems, and UTSA, etc. He received 2010 ACM Service Award and 2016 UTSA Innovation Award in the first category and 2014 Research Achievement Award from College of Science, UTSA. He is an associate editor of the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, the ACM Transactions on Multimedia Computing, Communications and Application, the Multimedia System Journal, and in the editorial board of the Journal of Multimedia and the Journal of Machine Vision and Applications. He is a fellow of the IEEE.