IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 29, 2019, accepted August 24, 2019, date of publication September 2, 2019, date of current version September 17, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2939071*

# Person Re-Identification With Joint Verification and Identification of Identity-Attribute Labels

**SHUN ZHANG[1], YANTAO HE[1], JIANG WEI[1], SHAOHUI MEI[1], (Member, IEEE), SHUAI WAN[1,2], (Member, IEEE), AND KE CHEN[3], (Senior Member, IEEE)**

[1]School of Electronic and Information, Northwestern Polytechnical University, Xi'an, China
[2]School of Engineering, Royal Melbourne Institute of Technology, Melbourne, VIC 3001, Australia
[3]Department of Computer Science, The University of Manchester, Manchester M13 9PL, U.K.

Corresponding author: Shun Zhang (szhang@nwpu.edu.cn)

This work was supported in part by the Youth Program of National Natural Science Foundation of China under Grant 61703344, in part by the Fundamental Research Funds for the Central Universities of China under Grant 3102017OQD021, and in part by the Top International University Visiting Program for Outstanding Young Scholars of Northwestern Polytechnical University.

**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. 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 supervised multi-task learning framework which considers attribute label information with joint identification-verification network to simultaneously learn an attribute-semantic and identity-discriminative feature representation. Specifically, this framework adopts the part-based deep neural network and learn three different tasks simultaneously: person identification, person verifications and attribute identification, so as to discover and capture concurrently complementary discriminative information about a person image from global and local image features and mid-level attribute features in one deep neural network. With the multi-task learning architecture, we obtain a discriminative model that reaches a synergy in distinguishing different person images, as manifested with the competitive accuracy on three person ReID datasets: Market1501, DukeMTMC-reID and VIPeR.

**INDEX TERMS** Person re-identification, attribute learning, multi-task learning, convolutional neural network.

## I. INTRODUCTION

Person re-identification (ReID) aims to identify a query person by finding the same persons among a set of gallery images or videos. It is drawing increasing research attentions in several computer vision applications including multi-camera tracking [1]–[4] and multi-camera activity analysis [5], smart city system [6], etc. The task is challenging owing to the dramatic changes in visual appearance from different camera views with non-overlapping visual fields resulting from variations in postures, view angles, occlusion, illumination, low resolution and background clutter [7]. Traditional researches on this topic mainly focus on two separate phases independently: either carefully feature designing (*i.e.* designing and extracting low-level visual features to represent person appearance) [8]–[11] or effective distance metric learning (*i.e.* learning a discriminative distance metric hence the distance of features from the same person can be smaller) [12]–[15].

The emergence of deep convolutional neural networks (CNNs) integrates the two separate phases mentioned above into a unified and end-to-end learning framework and therefore, introduces more powerful and robust feature representations. The CNN based approaches for ReID can be broadly classified into two categories: identification mode and verification mode. The identification mode utilizes a classification loss function (*e.g.* cross entropy loss) to learn a mapping function from raw images to person identity directly. This mode could make full use of the label information of the dataset, but it is prone to overfitting due to the limited

The associate editor coordinating the review of this article and approving it for publication was Mingjun Dai.

training samples for each person (*e.g.* , VIPeR [16] provides only 2 images per person). The verification mode includes the Siamese network with the contrastive loss [17], [18], and the triplet network with the triplet loss and its variants [19]–[22]. It learns a mapping function from raw images to the feature embedding space so that inter-class distance is enlarged and intra-class distance is decreased. As it considers a weak label (similar or not) between pairs, a drawback of the verification mode is not making full use of ReID label information.

Apart from issues stated above, an identification mode generally learns a mapping function from raw image pixels to high-level identities/classes directly, and the embedding space of the verification mode is hard to understand for humans. In contrast, human attributes (such as as hair-style, shoe-type or clothing-style) represent mid-level semantic description, *i.e*. a basis set defined by domain-specific axes which are semantically meaningful to humans, preferably unambiguous and more robust to subjective interpretation. Attributes are more consistent for the same person and more robust to the appearance variances in postures, view angles and so on, which can form a suitable representation for simply human understanding and direct human interaction. As shown in Figure 1, a ReID system may fail to discriminate between the first two persons in the first row wearing similar gray clothes and backpacks; but attributes may suggest that male or female, wearing a hat or not, dark or light shoes can eliminate the false matches.


**FIGURE 1. Examples of person images and attribute labels. Each pair of images represents the same person.**

Although some previous CNN-based person ReID approaches independently exploit joint identity and attribute label in model learning [23] or joint verification and identification [24]–[26], no one explicitly claims that the joint verification and identification supervision of both identity and attribute labels in one deep CNN model is critical to learn discriminative features for person re-identification. Based on the above considerations, in order to optimise person ReID task under significant viewing condition changes across cameras, we propose a joint identity-attribute learning approach with simultaneous verification and identification supervision in a unified deep learning model to discover and capture concurrently complementary discriminative information about a person image from both global and local features and mid-level semantic attributes. Specifically, we exploit the interaction and compatibility between mid-level semantic attributes and identity labels and discriminatively optimise both attribute and identity learning. In addition, we employ the joint supervision of verification and identification together and emphasize different aspects in feature learning: the identification supervisory signal tends to pull apart the deep network features of different identities since they have to be classified into different classes; the face verification signal requires that every two deep network feature vectors extracted from the same identity are close to each other while those extracted from different identities are kept away.

In summary, our contributions are three-fold: (1) we propose a new multi-task deep learning architecture which considers mid-level semantic attribute information by taking advantage of both multi-class identification and binary verification supervision signal to simultaneously learn an attribute-semantic and identity-discriminative feature representation more effectively; (2) we systematically analyze the relationship between person attribute and identity labels, and further exploit the interaction of verification and identification supervision; (3) we compare our approach with several state-of-art methods on three popular ReId datasets: Market1501 [27], DukeMTMC-reID [28] and VIPeR [29].

The rest of the paper is organized as follows: Section II reviews the related works with this paper. In Section III we describe the proposed network architecture and training algorithm. Section IV introduces the experiment results on three popular datasets. Last section draws the conclusion.

## II. RELATED WORK

Person re-identification has been a popular topic in computer vision and pattern recognition communities for a decade. There are several surveys [30], [31] on person re-identification. Traditional person re-identification works are accomplished by feature design and distance metric learning. In more recent years, the CNN based approaches for ReID have become a popular feature-based paradigm for learning discriminative feature representation. In this section, we are going to review recent ReID approaches based on deep learning and human semantic attributes and make a connection to our work.

### A. CNN-BASED APPROACHES

Generally speaking, two types of CNN models are commonly employed in the community: the identification mode and the verification mode. The identification mode could make full use of the ReID label information. Xiao *et al*. [32] present a domain guided dropout algorithm to improve the feature learning procedure from multiple domains and employ a softmax loss in the classification network. Zheng *et al*. [33] employ the identification mode to learn a discriminative embedding in the person subspace, and the work achieves good performance on the presented video ReID dataset.

The identification mode depends on lots of labeled training data per identity which might not be feasible for real-world deployment. In contrast, the verification mode considers the ReID task as a matching problem which could generate a good deal of sample pairs and would address the lack of

example issue. Some works employ the Siamese network with the contrastive loss. Ahmed *et al*. [34] improve the Siamese network by computing the cross-input neighborhood differences, which compares the features from one input image to features in neighboring locations of the other image. In [35], Varior *et al*. incorporate long short-term memory (LSTM) modules into the Siamese network to enhance the discriminative ability of the deep features. Some other works employ the Triplet network with the triplet loss and its variants. Su *et al*. [19] propose a three-stage learning process which includes attribute prediction using an independent dataset and an attributes triplet loss trained on datasets with ID labels. Hermans *et al*. [36] propose to generate the triplet with the hardest positive and hardest negative for each anchor sample in the mini-batch.

Some works adopt joint identification and verification supervision to train CNN in the field of face recognition and vehicle re-identification. In [24], Sun *et al*. present to learn the ''DeepID2'' features by the joint verification and identification supervision. Its extension works [25], [26] are also learned with the identification-verification supervisory signals. Some works [37]–[39] combine the verification and identification supervision to solve person ReID. For example, Chen *et al*. [39] employ classification loss and verification loss and develop a two-stepped fine-tuning strategy to transfer knowledge from auxiliary datasets.

Our network differs from the above in two aspects. First, we include attribute label information in joint identification-verification network to make full use of the label information for learning discriminative features more effective. Second, we leverage multi-task learning to exploit the interaction and compatibility between mid-level semantic attributes and identity labels and discriminatively optimise both attribute and identity learning.

### B. HUMAN SEMANTIC ATTRIBUTES

Human semantic attributes refer to semantic high level information which is invariant across many instances of person and contain information which is often highly relevant to a person's identity. Currently, many studies have applied deep learning to attributes learning for attribute prediction [40], [41] and person ReID [42]. In person re-identification, attributes show promising performance in preserving consistent representations of the same person and identifying differences among different persons [43]–[46]. Su *et al*. [47], [48] propose multi-task learning with low rank attribute embedding to address the problem of person re-identification on multi-cameras. Schumann and Stiefelhagen [49] present a person ReID approach which includes automatically predicted attribute information into the training process of a CNN. However, consistently predicting all attributes specific to a person is a difficult task when the labeled training data are sparse and person images in most person ReID datasets have low quality; *i.e*. inter-attribute discrimination can be weak on typical person ReID images.

Some person ReID approaches [23], [50], [51] employ both identity and attribute labels to improve re-identification performance. For example, Wang *et al*. [23] introduce a transferable joint deep learning by making full use of the label information of identities and attributes. Unlike their work that only utilizes the identification mode to learn joint attribute-identity feature, our approach further explores the verification mode to simultaneously learn an attribute-semantic and identity-discriminative feature representation in a multi-task learning framework.

## III. PROPOSED METHOD

In this section, we present our proposed person ReID approach. Firstly, we formulate the problem statement. Then we present our multi-task learning model for person re-identification with illustration. Finally we describe the loss function for learning with joint verification and identification of identity-attribute labels.

### A. PROBLEM FORMULATION

Assume a set of training images $\mathcal{I} = \{(\boldsymbol{I}_i, y_i, \boldsymbol{a}_i\}_{i=1}^{N}$ consisting of $N$ person bounding box images $\boldsymbol{I}_i$ with the corresponding identity labels as $y_i \in \mathcal{Y} = \{1, \ldots, C\}$ and attribute labels $\boldsymbol{a}_i \in \mathbb{R}^m$. These training images capture the visual appearance of $C$ different people under non-overlapping camera views. We define $\boldsymbol{a}_i = [a_{i,1}, \ldots, a_{i,m}]$ as an attribute label containing $m$ attributes, where $a_{i,j} \in \{0, 1\}$ is the binary indicator of the $j^{th}$ attribute. The **objective** is to formulate a deep learning model with the joint verification and identification supervision of attribute-identity labels to discover and capture concurrently complementary discriminative information about a person image from both global and mid-level visual features of the image.

### B. NETWORK ARCHITECTURE

Motivated by the previous works that extract discriminative features from different body parts of each person, we also construct a part-based deep CNN to extract global and local image features. Furthermore, our CNN model includes semantic attribute information with a multi-task learning architecture to discover complementary discriminative representation.

Figure 2 illustrates the architecture of our proposed framework. We adopt the ResNet-50 model, which has been evaluated to achieve competitive performances in some ReID systems, as the backbone of our network. The ResNet-50 model contains 5 different convolutional blocks: res_conv1, res_conv2_x (for x = 1,2,3), res_conv3_x (for x = 1,2,3,4), res_conv4_x (for x = 1,2,. . .,6) and res_conv5_x (for x = 1,2,3), where x indicates the number of blocks stacked (we add a name prefix ''res_'' to all blocks and please see Table 1 in the paper [52] for more details). The parts before res_conv4_2 block are the same with the original version of ResNet-50. The parts after res_conv4_1 block are divided into two branches: the global branch and the part branch. Both branches of the CNN model are trained to

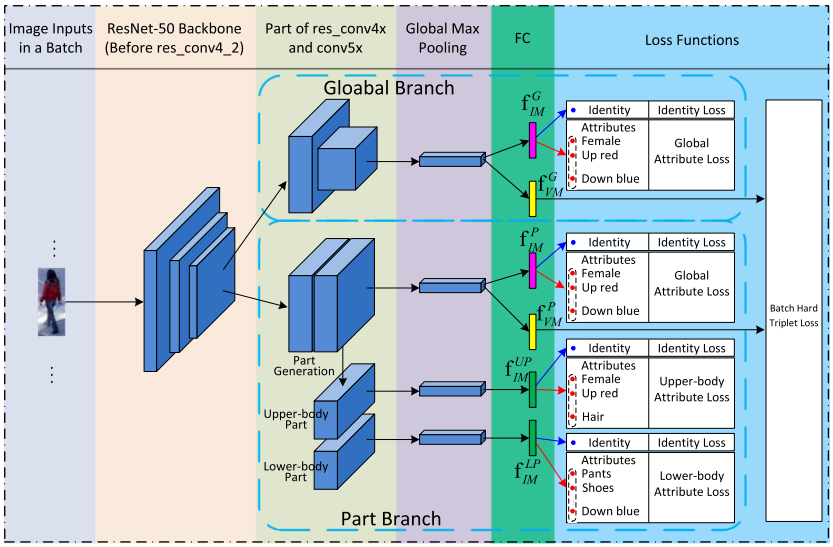



**FIGURE 2.** Our proposed multi-task learning architecture for person re-identification.

learn three different tasks simultaneously, including person identification, attribute identification and person verification.

In the global branch, parts of res_conv4_x and res_conv5_x have the same parameter settings with the original ResNet-50 model. The res_conv5_x block is connected with a global max-pooling (GMP) operation on the corresponding output feature map. The global branch outputs two independent fully connected features: $\boldsymbol{f}_{IM}^{G}$ is for joint person identification and attribute identification of global part of the input image, and $\boldsymbol{f}_{VM}^{G}$ is for person verification. This global branch learns the global feature representations and mid-level attributes without any partition information.

The part branch shares the similar network architecture with the global branch. The difference is that we employ no down-sampling operations in res_conv5_1 block to preserve proper areas of reception fields for local features of images, and we use a part generation strategy on the res_conv5_1 block: the output feature maps are split into two half parts: upper-body part and lower-body part, so that some attributes are more precisely located as shown in 1. After the GMP operation, the fully connected features $\boldsymbol{f}_{IM}^{P}$ and $\boldsymbol{f}_{VM}^{P}$ are for person/attribute identification and person verification, respectively. The two half parts are to learn local feature representations for person identification and attribute identification, in which $\boldsymbol{f}_{IM}^{UP}$ learns for upper-body attribute identification and $\boldsymbol{f}_{IM}^{LP}$ learns for lower-body attribute identification.

The network is trained end-to-end using the total loss:

$$L_{total} = L_{id} + \lambda_1 L_{attr} + \lambda_2 L_{trip-bh}, \tag{1}$$

where $L_{id}$ and $L_{attr}$ represent the person identification loss and attribute identification loss respectively, $L_{trip-bh}$ is the person verification loss with a triplet loss, and $\lambda_1$ and $\lambda_2$ balance the contribution of the three losses.

During testing phases, all the fully connected features ($\boldsymbol{f}_{IM}^{G}$, $\boldsymbol{f}_{VM}^{G}$, $\boldsymbol{f}_{IM}^{P}$, $\boldsymbol{f}_{VM}^{P}$, $\boldsymbol{f}_{IM}^{UP}$ and $\boldsymbol{f}_{IM}^{LP}$) are concatenated as the final feature in order to extract both the global and local information to obtain the most powerful discrimination of learned features.

### C. LOSS FUNCTIONS

In this subsection, we present our loss functions employed in our multi-task learning architecture via the exploitation of Person identification loss, attribute identification loss and batch hard triplet loss.

#### 1) PERSON IDENTIFICATION LOSS

For the person identification loss we utilize the cross-entropy classification loss function by optimising person identification task given training labels of multiple person identities. Formally, we predict the prediction probability of image $I_i$ over the given identity label $y_i$ as:

$$p_{id}(\boldsymbol{I}_i, y_i) = \frac{\exp(\boldsymbol{w}_{y_i}^T \boldsymbol{f}_{IM}(\boldsymbol{I}_i))}{\sum_{k=1}^{C} \exp(\boldsymbol{w}_k^T \boldsymbol{f}_{IM}(\boldsymbol{I}_i))}, \tag{2}$$

where $\boldsymbol{f}_{IM}(\boldsymbol{I}_i)$ refers to the feature vector ($\boldsymbol{f}_{IM}^{G}$ or $\boldsymbol{f}_{IM}^{P}$ in Fig. 2) of $\boldsymbol{I}_i$ and $\boldsymbol{w}_k$ the prediction function parameter of training identity class $k$. The training loss on a batch of $n_{bs}$ images is computed as:

$$L_{id} = -\frac{1}{n_{bs}} \sum_{i=1}^{n_{bs}} \log(p_{id}(\boldsymbol{I}_i, y_i)). \tag{3}$$

This identification loss encourages the network to learn more discriminative identity-related features (*i.e.* , features with large inter-personal variations), which can help correctly classify all the classes simultaneously.

**TABLE 1. Attributes distribution of global part, upper-body part and lower-body part.**

| Global Part | |
|---|---|
| Upper-body Part | Lower-body Part |
| age, gender, hair length, sleeve length, carrying bag, carrying backpack, colors of upper-body clothing, wearing hat, carrying handbag | length of lower-body, type of lower-body clothing, colors of lower-body clothing, shoe type, color of shoes |

#### 2) ATTRIBUTE IDENTIFICATION LOSS

The attribute identification loss includes three terms: global attribute loss $L^G_{attr}$, upper-body attribute loss $L^{UP}_{attr}$ and lower-body attribute loss $L^{LP}_{attr}$:

$$L_{attr} = L^G_{attr} + L^{UP}_{attr} + L^{LP}_{attr}. \tag{4}$$

The problem of attribute prediction is treated as multi-label classification. We pass the last fully connected layers into a sigmoid layer to squash the predicted classification scores to [0, 1]. The global attribute prediction of $L^G_{attr}$ is optimized using binary entropy-cross loss function by considering all $m$ attribute classes:

$$L^G_{attr} = -\frac{1}{n_{bs}} \sum_{i=1}^{n_{bs}} \sum_{j=1}^{m} (a_{i,j} \log(p_{att}(\boldsymbol{I}_i, a_{i,j})) \tag{5}$$

$$+ (1 - a_{i,j}) \log(1 - p_{att}(\boldsymbol{I}_i, a_{i,j})), \tag{6}$$

where $a_{i,j}$ and $p_{att}(\boldsymbol{I}_i, a_{i,j})$ define the groundtruth attribute label and the predicted classification probability on the $j^{th}$ attribute class of the training image $\boldsymbol{I}_i$. Similarly, the upper-body attribute loss $L^{UP}_{attr}$ and lower-body attribute loss $L^{LP}_{attr}$ are calculated with the groundtruth attribute labels (in Table 1) and the predicted classification probabilities.

#### 3) BATCH HARD TRIPLET LOSS

Traditional triplet Loss takes a triplet as input. One triplet consists of an anchor sample $\boldsymbol{I}_{s^o}$, a positive sample $\boldsymbol{I}_{s^+}$ which belongs to the same identity with the anchor sample and a negative sample $\boldsymbol{I}_{s^-}$ which has a different identity with the anchor sample. The loss aims to ensure that the embedded distance $D_{s^o,s^+}$ between the anchor sample $\boldsymbol{I}_{s^o}$ and positive sample $\boldsymbol{I}_{s^+}$ is closer than the distance $D_{s^o,s^-}$ of the anchor sample $\boldsymbol{I}_{s^o}$ and negative sample $\boldsymbol{I}_{s^-}$ by a distance margin $\alpha$. For all triplets in one mini-batch, the triplet loss is of the form:

$$\mathcal{L}_{trip} = \sum_{\substack{s^o,s^+,s^- \\ y_{s^o}=y_{s^+}\neq y_{s^-}}} [D_{s^o,s^+} - D_{s^o,s^-} + \alpha]_+, \tag{7}$$

where the subscript plus sign of square brackets is defined as $[x]_+ = \max\{0, x\}$.

The traditional Triplet Loss needs to sample the training dataset and generate training triplets at the beginning of the training. Since the generated triplets depend on the initial CNN model, it will not be corrected after the training proceeds. The core idea of the Batch Hard Triplet Loss is to combine the triplet generation step with the training process and to mine the hard triplet samples within each mini-batch. The mini-batch is organized as follows: each mini-batch is filled by randomly sampling $P$ persons and then $K$ instances was sampled for each identity, *i.e.* each min-batch is filled with $P \times K$ images. Each image in the mini-batch is in turn treated as an anchor sample. Batch Hard Triplet Loss tries to choose the hardest positive and negative sample in the mini-batch:

$$\mathcal{L}_{trip-bh} = \sum_{i=1}^{P} \sum_{s^o=1}^{K} [D^*_{s^o,s^+} - D^*_{s^o,s^-} + \alpha]_+, \tag{8}$$

where $D^*_{s^o,s^+}$ and $D^*_{s^o,s^-}$ denote the hardest positive and negative sample corresponding to anchor sample $\boldsymbol{I}_{s^o}$ as:

$$D^*_{s^o,s^+} = \sum_{i=1}^{P} \sum_{s^o=1}^{K} \max_{s^+=1,\ldots,K} D(\boldsymbol{f}_{VM}(\boldsymbol{I}^i_{s^o}), \boldsymbol{f}_{VM}(\boldsymbol{I}^i_{s^+})), \tag{9}$$

$$D^*_{s^o,s^-} = \sum_{i=1}^{P} \sum_{s^o=1}^{K} \min_{\substack{s^-=1,\ldots,K \\ j=1,\ldots,P \\ j\neq i}} D(\boldsymbol{f}_{VM}(\boldsymbol{I}^i_{s^o}), \boldsymbol{f}_{VM}(\boldsymbol{I}^j_{s^-})), \tag{10}$$

where $\boldsymbol{f}_{VM}(\boldsymbol{I}_i)$ refers to the feature vector ($\boldsymbol{f}^G_{VM}$ or $\boldsymbol{f}^P_{VM}$ in Fig. 2) of $\boldsymbol{I}_i$.

## IV. EXPERIMENTS

In this section, we report the experimental results on three popular and relatively large benchmark datasets: Market1501 [27], DukeMTMC-reID [28] and VIPeR [29]. First, we introduce the implementation details and dataset information. Then we compare our approach against several state-of-the-art methods. We further conduct an ablation study to demonstrate the contribution of individual modules in our approach. Finally, we present some examples of ranking and attribute recognition results.

### A. IMPLEMENTATION DETAILS

In our experiments, the parameters of ResNet-50 model are pre-trained on the ImageNet dataset. During the training and testing phase, input images are resized to $384 \times 128$ with normalization using the standard deviation and mean. The batch-size is set to 32, with 32 different identities and 4 instances per identity in each mini-batch. We used Stochastic Gradient Descent (SGD) algorithm and the training process takes 320 epochs in total. The initial learning rate is set to $2 \times 10^{-4}$ and decayed at the 160th epoch according to the "exp" decay rule of Pytorch. The margin $\alpha$ is set to 1.2 empirically. In addition, we use Re-ranking [53] as the post-processing step to get better performance. The proposed architecture is implemented using the Pytorch [54] deep learning framework and two NVIDIA Titan GPU, which consists of 3584 CUDA cores, 10 Gbps memory speed, and 12 GB GDDR5X of standard memory configuration.

**TABLE 2. The specifications of the two evaluated ReID datasets.**

| Datasets | VIPeR | Market-1501 | DukeMTMC-reID |
|---|---|---|---|
| #identities | 632 | 1,501 | 1,812 |
| #images | 1,264 | 32,643 | 36,411 |
| #cameras | 2 | 6 | 8 |
| #training IDs | 316 | 750 | 702 |
| #test IDs | 316 | 751 | 702 |
| #probe images | 316 | 3,368 | 2,228 |
| #gallery images | 316 | 19,732 | 17,661 |
| #attributes | 19 | 27 | 23 |

### B. DATASETS

We evaluate our approach on three large-scale public datasets. Table 2 summarizes the statistics of three datasets, including the number of training and test identities, probe and gallery images and attributes.

#### 1) Market1501

The Market1501 [27] dataset contains 32,688 annotated bounding boxes of 1,501 identities in total. A total of six cameras are used to collect images. There are 12,936 images of 751 identities in the training set and 19,732 images of 750 identities in the testing set including 3,368 query images and 19,732 gallery images. All the images are detected by the Deformable Part Model instead of hand drawing. Thus it is closer to the practical applications. The Market-1501 dataset is annotated with 27 attributes such as gender, hair length, age, carrying bag, sleeve length, and colors of upper-body clothing and lower-body clothing. The training and test sets are pre-split by the contributors of [27]. The test set is evaluated with the Single-Query setting which means we select only one of the query images each time and then find the same person with the query image among the set of gallery images.

#### 2) DukeMTMC ReID

The DukeMTMC-reID [28] dataset is a subset of the DukeMTMC for image-based re-identification. The original dataset contains 85-minute high-resolution videos from 8 different cameras, and provides manually labeled bounding boxes. It has the same form as Market1501. There are 36,411 bounding boxes in total, with 1,404 identities appearing in more than two cameras and 408 identities (distractor ID) appearing only in one camera. The training set contains 702 IDs and the testing set has 702 IDs. In the testing set, one query image for each ID in each camera is selected and the remaining images are put into the gallery. Eventually, it contains 16,522 training images from 702 people and 2,228 query images from another 702 people, and a search gallery of 17,661 images. There are 23 classes of attributes labelled for the DukeMTMC-reID dataset, such as shoes type and color, wearing hat, carrying bag, carrying backpack, carrying handbag. The training and test sets are pre-split by the contributors of [28]. The test set is also evaluated with the Single-Query setting.

#### 3) VIPeR

VIPeR [29] is featured with low resolution. Although it has been tested by many researchers, it's still one of the most challenging datasets. It contains 632 identities and each has two images captured from two viewpoints with distinct view angles under varying illumination conditions. The images are normalized to $128 \times 48$ pixels. The attributes of VIPeR are annotated in the PETA [55] dataset. Each image is labeled with 61 binary and 4 multi-class attributes, and we select 19 attributes that have more than 50 training images, such as gender, hair length, carrying backpack, upper-body Tshirt, lower-body Jeans, Shoes. All images are randomly divided into two equal halves: one for training and the other for testing. This is repeated for 10 times and the averaged performance is reported. Here, we use the VIPeR data split provided by [56]. The test set is evaluated with the Single-Query setting.

#### 4) EVALUATION METRICS

For the person ReID task, the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) are used for evaluation. The CMC curve shows the probability that a query image appears in different-sized candidate lists. For each query image, an algorithm will rank all the gallery images according to their distances to the query image from small to large, and the CMC Top-k accuracy is: $Acc_k = 1$ if Top-$k$ ranked gallery images contain the query identity, otherwise $Acc_k = 0$. The final CMC curve is computed by averaging $Acc_k$ over all the queries [57]. This measurement means people care more about returning the ground truth match in the top positions of the rank list. The mAP metric is used to evaluate the overall performance. It is to measure the recall of multiple ground truth matches, computed by first computing the area under the Precision-Recall curve for each query image, then calculating the mean of Average Precision over all query images [27]. In the experiments, we use the evaluation package publicly available in [27], [58].

For the attribute recognition task, we test the classification accuracy for each attribute. The gallery images are used as the testing set. When testing the attribute prediction, we omit the distractor (background) and junks images, since they do not have attribute labels. We report the independent accuracy of all these attribute predictions and the averaged accuracy as the overall attribute prediction accuracy.

### C. COMPARATIVE STUDY

We compare our results against the state-of-the-art methods on the Market-1501, DukeMTMC-reID and VIPeR datasets with Single-Query setting. As shown in Tables 3 to 5, we compare the proposed approach with a set of related, top performing deep learning based ReID approaches. We divide the related approaches into three parts: Verification Mode (VM), Identification Mode (IM) and combination of Identification Mode and Verification Mode (IM+VM). We indicate the used labels of each work for training: "ID"

**TABLE 3.** Experimental results(%) of the presented approach and other comparisons on the Market-1501 dataset with single-query setting. The CMC Top-1-5-10 and mAP accuracies are reported. The "IM" and "VM" denote identification mode and verification mode, respectively, and the "ID" and "Attr" denote using identity label and attribute label, respectively. The "woRK" means without using the Re-ranking [53] algorithm. The accuracies of the best performing approaches are marked in bold.

| Methods | Network | mode | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|---|---|
| MultiRegion [59] | Designed net | VM+ID | 66.4 | 85.0 | 90.2 | 41.2 |
| DeepPartAR [60] | GoogLeNet | VM+ID | 81.0 | 92.0 | 94.7 | 63.4 |
| TriNet [36] | ResNet-50 | VM+ID | 84.9 | 94.2 | - | 69.1 |
| JLMultiLoss [61] | ResNet-50 | IM+ID | 85.1 | - | - | 65.5 |
| CStyle [62] | ResNet-50 | IM+ID | 89.5 | - | - | 71.6 |
| PartLoss [63] | GoogLeNet | IM+ID | 88.2 | - | - | 69.3 |
| DPFL [64] | Inception-V3 | IM+ID | 88.9 | - | - | 73.1 |
| GLAD [65] | GoogLeNet | IM+ID | 89.9 | - | - | 73.9 |
| HA-CNN [66] | Designed net | IM+ID | 91.2 | - | - | 75.7 |
| PCB-RPP [67] | ResNet-50 | IM+ID | 93.8 | - | - | 81.6 |
| EBB [68] | Designed net | IM+ID | 81.2 | 94.6 | 97.0 | - |
| DSR [69] | ResNet-50 | IM+ID | 83.6 | - | - | 64.3 |
| HydraPlus [50] | Designed net | IM+ID+Attr | 76.9 | 91.3 | 94.5 | - |
| APR [51] | ResNet-50 | IM+ID+Attr | 87.0 | 95.1 | 96.4 | 66.9 |
| AACN [37] | GoogLeNet | IM+VM+ID | 88.7 | - | - | 83.0 |
| DLCE [38] | ResNet-50 | IM+VM+ID | 79.5 | - | - | 59.9 |
| DTL [39] | GoogLeNet | IM+VM+ID | 83.7 | - | - | 65.6 |
| Ours(woRK) | ResNet-50 | IM+VM+ID+Attr | 92.7 | 96.8 | **98.1** | 81.5 |
| Ours | ResNet-50 | IM+VM+ID+Attr | **94.7** | **97.0** | 97.9 | **88.4** |

**TABLE 4.** Experimental results(%) of the presented approach and other comparisons on the Duke-MTMC ReID dataset. The CMC Top-1 and mAP accuracies are reported. The "woRK" means without using the Re-ranking [53] algorithm. The accuracies of the best performing approaches are marked in bold.

| Methods | Network | mode | Top-1 | mAP |
|---|---|---|---|---|
| CStyle [62] | ResNet-50 | IM+ID | 78.3 | 57.6 |
| DPFL [64] | Inception-V3 | IM+ID | 79.2 | 60.6 |
| HA-CNN [66] | Designed net | IM+ID | 80.5 | 63.8 |
| PCB-RPP [67] | ResNet-50 | IM+ID | 82.9 | 68.5 |
| APR [51] | ResNet-50 | IM+ID+Attr | 73.9 | 55.6 |
| Ours(woRK) | ResNet-50 | IM+VM+ID+Attr | 83.1 | 80.2 |
| Ours | ResNet-50 | IM+VM+ID+Attr | **85.6** | **85.8** |

and "Attr" denote using identity labels and attribute labels, respectively. We also include the used network backbone of each approach in the table. Most methods such as TriNet, CStyle, PCB-RPP, DSR and APR use the similar backbone network of ResNet-50 with our network.

On the Market-1501 dataset, APR (IM+ID+Attr) uses identification mode and employs both identity and attribute labels, and it achieves 87.0% Top-1, 95.1% Top-5, 96.4% Top-10 and 66.9% mAP. AACN (IM+VM+ID) combines identification mode and verification mode, and achieves 88.7% Top-1 and 83.0% mAP. In contrast, Our work (IM+VM+ID+Attr) combines the Identification Mode and Verification Mode with both identity and attribute labels, and Ours(woRK) increases Top-1 to 92.7%, Top-5 to 96.8%, Top-10 to 98.1%, and mAP to 81.5%. Furthermore, combined with the Re-ranking algorithm, our method obtains the best accuracy on Top-1 94.7%, Top-5 97.0%, and mAP 88.4%.

On DukeMTMC-reID, PCB-RPP (IM+ID) achieves 68.5% mAP and 82.9% for Top-1 accuracy, and APR (IM+ID+Attr) obtains 55.6% mAP an 73.9% Top-1. Our method (IM+VM+ID+Attr) without using Re-ranking increases mAP to 80.2% and Top-1 to 83.1%. Our full model with Re-ranking obtains the best accuracy on mAP 85.8% and Top-1 85.6%. The results show that our method could discover more discriminative representation from both image features and mid-level attributes with joint identification and verification supervision.

**TABLE 5.** Experimental results(%) of the presented approach and other comparisons on the VIPeR [29] dataset. The CMC Top-1-5-10 and mAP accuracies are reported. The accuracies of the best performing approaches are marked in bold.

| Methods | Network | mode | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|---|---|
| Quadruplet [70] | Designed net | VM+ID | 49.1 | 73.1 | 82.0 | - |
| SiaLSTM [35] | Designed net | VM+ID | 42.4 | 68.7 | 79.4 | 47.9 |
| Gated S-CNN [71] | Designed net | VM+ID | 37.8 | 66.9 | 77.4 | - |
| DeepRank [72] | AlexNet | VM+ID | 38.4 | 69.2 | 81.3 | - |
| ImpTrpLoss [21] | Designed net | VM+ID | 47.8 | 74.4 | 84.8 | - |
| JLMultiLoss [61] | ResNet-50 | IM+ID | 50.2 | 74.2 | 84.3 | - |
| PCB-RPP [67] | ResNet-50 | IM+ID | 38.1 | 53.2 | 59.3 | 45.4 |
| PPA [73] | Resnet-50 | IM+ID | 45.1 | 65.1 | 72.7 | 54.5 |
| MTDnet [74] | AlexNet | IM+VM+ID | 47.5 | 73.1 | 82.6 | - |
| AlignedReID [75] | Resnet-50 | IM+VM+ID | 42.8 | 63.7 | 73.6 | 53.0 |
| Ours | ResNet-50 | IM+VM+ID+Attr | **51.6** | **76.0** | **87.0** | **57.3** |

On the VIPeR dataset, JLMultiLoss (IM+ID) adopts the identification mode with the person identity information, and it achieves 50.2% Top-1, 74.2% Top-5 and 84.3% Top-10. AlignedReID (IM+VM+ID) combines the identification mode and verification mode to address the person ReID problem and achieves 42.8% Top-1, 63.7% Top-5, 73.6% Top-10 and 53.0% mAP. In contrast, our method (IM+VM+ID+Attr) obtains the best performance on all metrics: Top-1 51.6%, Top-5 76.0%, Top-10 87.0% and mAP 57.3%, as our method takes advantage of both multi-class identification and binary verification supervision signals to simultaneously learn an attribute-semantic and identity-discriminative features.

### D. RESULTS ON PERSON ATTRIBUTE RECOGNITION

We test person attribute recognition on the galleries of the Market-1501 and DukeMTMC-reID datasets in Table 6 and Table 7, respectively.

By comparing the results of the state-of-the-art approaches, APR and ARN in [51], the proposed approach achieves the best attribute recognition accuracy in overall. The improvements in average are 1.15% and 5.40% on Market-1501 and DukeMTMC-reID, respectively. Our approach performs favorably against the other methods in the recognition of color of upper-body clothing and color of lower-body clothing: on the Market-1501 dataset, the improvements are 27.84% and 28.3%, respectively; on the DukeMTMC-reID dataset, the improvements are 33.48% and 24.26%, respectively. In our approach, the integration of identity recognition and identity verification introduce some complementary

**TABLE 6.** Attribute recognition accuracy on the Market-1501 dataset. "L.slv", "L.low", "S.clth", "C.up" and "C.low" denote length of sleeve, length of lower-body clothing, style of clothing, color of upper-body clothing and color of lower-body clothing, respectively. The best values are highlighted in bold.

| Methods | gender | age | hair | L.slv | L.low | S.clth | backpack | handbag | bag | hat | C.up | C.low | average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARN [51] | 87.5 | 85.8 | 84.2 | 93.5 | 93.6 | **93.6** | 86.6 | 88.1 | 78.6 | 97.0 | 72.4 | 71.7 | 86.0 |
| APR [51] | **88.9** | 88.6 | 84.4 | **93.6** | **93.7** | 92.8 | 84.9 | 90.4 | 76.4 | 97.1 | 74.0 | 73.8 | 86.6 |
| Ours | 84.5 | **91.6** | **84.9** | 78.3 | 84.7 | 80.9 | **86.7** | **92.3** | **80.2** | **97.6** | **94.6** | **94.7** | **87.6** |

**TABLE 7.** Attribute recognition accuracy on the DukeMTMC-reID dataset. "L.up", "C.shoes", "C.up" and "C.low" denote length of sleeve, color of shoes, color of upper-body clothing and color of lower-body clothing, respectively. The best values are highlighted in bold.

| Methods | gender | hat | boots | L.up | backpack | handbag | bag | C.shoes | C.up | C.low | average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ARN [51] | 82.0 | 85.5 | 88.3 | 86.2 | **77.5** | 92.3 | 82.2 | 87.6 | 73.4 | 68.3 | 82.3 |
| APR [51] | 84.2 | 87.6 | 87.5 | **88.4** | 75.8 | 93.4 | **82.9** | 89.7 | 74.2 | 69.9 | 83.4 |
| Ours | **85.4** | **89.3** | **88.6** | 86.6 | 76.7 | **93.6** | 82.0 | **91.6** | **92.2** | **93.3** | **87.9** |

information which is helpful for learning a more discriminative attribute model.

We also observe that the recognition rate of some attributes decreases for our approach, such as length of lower-body clothing and style of clothing in Market-1501, and backpack and bag in DukeMTMC-reID. The reason is that our model is optimized for re-ID, some ambiguous images of certain attributes may be incorrectly predicted. Nevertheless, the improvement on the two datasets is still encouraging.

### *E. ABLATION STUDY*

Our person re-identification approach learns the discriminative features with the joint verification and identification of person labels and attributes. We conduct an ablation study to demonstrate the effectiveness of individual modules in our approach by removing modules.

1) Ours-woID: without person identification loss $L_{id}$;
2) Ours-woAttr: without attribute identification loss $L_{attr}$;
3) Ours-woIM: without identification mode, *i.e.* without $L_{attr}$ and $L_{id}$;
4) Ours-woVM: without verification mode, *i.e.* without batch hard triplet loss $L_{trip-bh}$.

To eliminate the effect of post-processing on accuracy, all experiments are conducted **without Re-ranking** on Market-1501 with the metrics of mAP and Top-1-5-10.

**Effectiveness of identity recognition.** To evaluate the effectiveness of identity recognition, we compare our model with Ours-woID model (without the supervision of identity labels) while keeping all other factors are the same. Table 8 shows the performance of Ours-woID decreases in all metric terms of Person ReID. For example, Top-1 and mAP of Ours-woID drop from 92.7% to 82.3% and 81.5% to 62.8%, respectively. The results show that with the help of joint identification of identity and attribute labels, the proposed method improves the re-identification accuracy.

**Effectiveness of attribute recognition.** To evaluate the effectiveness of attribute recognition, we compare our full model with Ours-woAttr (without the supervision of attribute labels) model. As shown in Table 8, Ours-woAttr model without attribute recognition has lower performance in all metric terms. Specifically, Ours-woAttr decreases mAP to 77.8% and Top-1 to 91.5%. The results show that our presented approach with the attribute recognition could improve the final person re-identification performance.

**TABLE 8.** Ablation study on the Market-1501 dataset. All experiments are conducted without Re-ranking. The best values are highlighted in bold.

| Methods | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|
| Ours-woID | 82.3 | 91.7 | 94.7 | 62.8 |
| Ours-woAttr | 91.5 | 96.7 | 98.0 | 77.8 |
| Ours-woIM | 80.1 | 91.5 | 94.6 | 61.0 |
| Ours-woVM | 88.7 | 95.7 | 97.2 | 71.4 |
| Ours | **92.7** | **96.8** | **98.1** | **81.5** |

**Effectiveness of identification mode.** To evaluate the Effectiveness of identification mode, Ours-woIM model just adopts the verification supervision, *i.e.* the loss function just employs the batch hard triplet loss function. Table 8 shows that without help of identification supervision signal, Ours-woIM model achieves the lowest performance in all metric terms. For example, the mAP of Ours-woIM drops to 61.0% and Top-1 to 80.1%, which shows the identification mode is critical for person re-identification performance.

**Effectiveness of verification mode.** To evaluate the effectiveness of verification mode, we compare our model with Ours-woVM model. Table 8 shows that without verification mode, the mAP of Ours-woVM model drops to 71.4%, which shows verification mode is also critical for person re-identification performance.

*Our findings are the following:* (i) Ours-woIM achieves the lowest performance in Table 8, which shows the Identification mode is the most important component for person ReID. The conclusion is consistent with the results in Table 3 where methods of IM (*e.g.* PCB-RPP) have higher mAP and Top-1 accuracy than those of VM (*e.g.* TriNet). (ii) The performance of our model goes down obviously when removing any individual modules. In other words, our methods of joint identity-attribute learning with simultaneous verification and identification supervision can improve person re-identification performance.

### F. QUALITATIVE RESULTS

We present sample ranking results in the Market-1501 dataset to show the performance of the method in effective retrieval, as well as show the false positive retrieval results as presented in Figure 3. It can be seen from the ranking results that the proposed method has a good retrieval ability and achieves a higher retrieval rate of true positive matches than that of false positives. We can also see that, our approach can retrieve persons with poses significantly different from poses of query images. For example, in the first row of Figure 3, although the give query image is the front of the person, our Top-10 ranking images include diversified poses of the same person, which shows the robustness of our model.

In addition, we present the qualitative results of attributes recognition performed on the attributes datasets in Fig. 4. We can see that the model suffers when predicting attributes related to objects carried by the pedestrian, such as bag and backpack, and are found to be challenging to infer correctly due to the occlusion by pedestrian.

**FIGURE 3.** Example query images and their ranking lists for our method on Market-1501. Images in the first left column are query images, and the right 10 images are ranking results. Green boundary is added to the true positive and red to false positive.

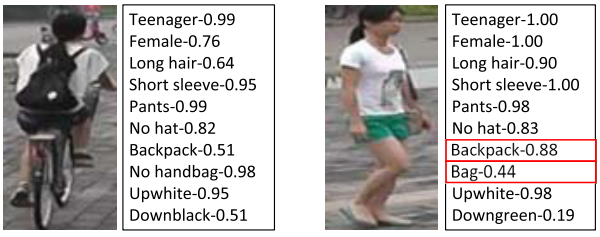



**FIGURE 4.** Examples for person attribute recognition. The two tables show the predicted attributes and the classification scores. Red bounding boxes indicate incorrect predictions.

## V. CONCLUSION

In this paper, we present a multi-task deep learning architecture that makes joint use of person identification, attribute identification and person verification in one unified deep learning network to learn discriminative features for person re-identification. The identification supervisory signal pulls apart the deep network features of different identities; the person verification signal requires that deep network feature vectors from the same identity are close to each other while those extracted from different identities are kept away; and the identity labels provide high-level classes of person images while human attributes provide mid-level semantic information. Through fusion of with the identification and verification supervision and complementary information of attribute and identity labels, our model achieves competitive accuracy to several state-of-the-art methods on two person ReID datasets. We will exploit the constraints between attributes (*e.g.* dress is more likely related to female) to improve person re-identification performance.

## ACKNOWLEDGMENT

The work was partially conducted when the first author visited in Department of Computer Science at University of Manchester.

## REFERENCES

[1] J. Berclaz, F. Fleuret, and P. Fua, "Multi-camera tracking and atypical motion detection with behavioral maps," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 112–125.
[2] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.
[3] S. Zhang, J. Wang, Z. Wang, Y. Gong, and Y. Liu, "Multi-target tracking by learning local-to-global trajectory models," *Pattern Recognit.*, vol. 48, no. 2, pp. 580–590, 2015.
[4] M. Amini-Omam, F. Torkamani-Azar, and S. A. Ghorashi, "Maximum likelihood estimation for multiple camera target tracking on Grassmann tangent subspace," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 77–89, Jan. 2018.
[5] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 106–129, 2010.
[6] P. Vitello, A. Capponi, C. Fiandrino, P. Giaccone, D. Kliazovich, and P. Bouvry, "High-precision design of pedestrian mobility for smart city simulators," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
[7] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia, "Person re-identification with triplet focal loss," *IEEE Access*, vol. 6, pp. 78092–78099, 2018.
[8] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. Brit. Machive Vis. Conf.*, 2012, p. 11.
[9] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 391–401.
[10] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.
[11] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 244–249.
[12] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3567–3574.
[13] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 501–512.
[14] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 1–16.
[15] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[16] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 262–275.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[18] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1983–1991.

[19] 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.* Berlin, Germany: Springer, 2016, pp. 475–491.

[20] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, Sep. 2016, p. 3.

[21] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[22] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3741–3750.

[23] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.

[24] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[25] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2892–2900.

[26] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," 2015, *arXiv:1502.00873*. [Online]. Available: https://arxiv.org/abs/1502.00873

[27] 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.*, Dec. 2015, pp. 1116–1124.

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 17–35.

[29] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. VS-PETS Workshop*, vol. 3, no. 5, 2007, pp. 1–7.

[30] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, 2014.

[31] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[32] 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.

[33] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 868–884.

[34] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3908–3916.

[35] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 135–153.

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

[37] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, " Attention-aware compositional network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.

[38] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, p. 13, 2017.

[39] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian, and T. Xiang, "Deep transfer learning for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Big Data*, Sep. 2018, pp. 1–5.

[40] S. Shankar, V. K. Garg, and R. Cipolla, "Deep-carving: Discovering visual attributes by carving deep neural nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3403–3412.

[41] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5315–5324.

[42] 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.*, Jun. 2015, pp. 4184–4193.

[43] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proc. BMVC*, 2012, vol. 2, no. 3, p. 8.

[44] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 402–412.

[45] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 93–117.

[46] R. Layne, T. M. Hospedales, and S. Gong, *Re-Id: Hunting Attributes in the Wild*. Durham, U.K.: BMVA Press, 2014.

[47] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-Task learning with low rank attribute embedding for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3739–3747.

[48] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for multi-camera person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1167–1181, May 2018.

[49] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 20–28.

[50] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 350–359.

[51] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2018.

[52] 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.

[53] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1318–1327.

[54] *PyTorch*. Accessed: Sep. 5, 2019. [Online]. Available: https://pytorch.org/

[55] 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.

[56] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.

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

[58] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3754–3762.

[59] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2017, pp. 1–6.

[60] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3219–3228.

[61] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2194–2200.

[62] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.

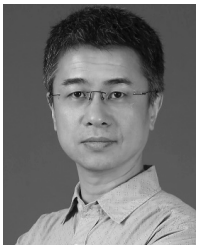
[63] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[64] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2590–2600.

[65] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2017, pp. 420–428.
[66] 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.
[67] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 480–496.
[68] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5794–5803.
[69] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7073–7082.
[70] 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.
[71] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 791–808.
[72] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.
[73] S. Qiao, C. Liu, W. Shen, and A. L. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7229–7238.
[74] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 3988–3994.
[75] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: https://arxiv.org/abs/1711.08184

**SHUN ZHANG** received the B.S. and the Ph.D. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2016, respectively. He is currently an Assistant Professor with the School of Electronic and Information, Northwestern Polytechnical University, Xi'an. His research interests include machine learning, computer vision, and human–computer interaction, with a focus on visual tracking, object detection, image classification, feature extraction, and sparse representation.

**YANTAO HE** received the B.S. degree in electronic information engineering and the M.E. degree in signal information processing from Northwestern Polytechnic University, China, in 2016 and 2019, respectively. His research interests include pattern recognition, object tracking, and person re-identification based on convolutional neural networks. He received the second-class scholarship of Northwest Polytechnic University for three consecutive years.



**JIANG WEI** received the B.S. degree in computer and applications from the Xi'an University of Architecture and Technology, Xi'an, China, in 1990, and the M.S. degree in circuits and systems from Northwestern Polytechnical University, Xi'an, where he has been an Associate Professor with the School of Electronics and Information, since 2003. His research interests include image processing, computer vision, and their applications in embedded systems.

**SHAOHUI MEI** received the B.S. degree in electronics and information engineering and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2011, respectively, where he is currently an Associated Professor with the School of Electronics and Information. He was a Visiting Student with The University of Sydney, from October 2007 to October 2008. His research interests include hyperspectral remote sensing image processing, deep learning, video processing, and pattern recognition.

**SHUAI WAN** (S'06–M'08) received the B.E. degree in telecommunication engineering and the M.E. degree in communication and information systems from Xidian University, China, in 2001 and 2004, respectively, and the Ph.D. degree in electronic engineering from Queen Mary University of London, in 2007. She joined Northwestern Polytechnical University, where she is currently a Professor. Since 2016, she has been an Adjunct Professor with the School of Engineering, Royal Melbourne Institute of Technology, Australia. Her research interests include versitile video coding, video processing, and streaming.

**KE CHEN** (M'97–SM'00) received the B.Sc. and M.Sc. degrees from Nanjing University, Nanjing, China, in 1984 and 1987, respectively, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1990, all in computer science.

He was with Peking University, The University of Birmingham, Ohio State University, Kyushu Institute of Technology, and Tsinghua University. He was also a Visiting Professor with The Hong Kong Polytechnic University and a Visiting Senior Researcher with Microsoft Research Asia. Since 2003, he has been with The University of Manchester, Manchester, U.K., where he directs the Machine Learning and Perception (MLP@UoM) Lab. He has published more than 200 articles in peer-reviewed journals and conferences. His current research interests include machine learning, machine perception, computational cognitive modeling, and their applications in intelligent system development.

Dr. Chen was a member of IEEE Biometrics Council AdCom, in 2012 and 2013. He received several academic awards, including the NSFC Distinguished Principal Young Investigator Award, in 2001, and the JSPS Research Award, in 1993. In 2008 and 2009, he chaired the IEEE Computational Intelligence Society's Intelligent Systems Applications Technical Committee and the IEEE Computational Intelligence Society's University Curricula Subcommittee. He was a technical program Co-Chair of several conferences, e.g., *International Joint Conference on Neural Networks* (IJCNN'12), and has been a member of the technical program committee of numerous international conferences. He has been an Associate Editor and on an editorial board of several academic journals, including *Neural Networks*, since 2011, and the IEEE TRANSACTIONS ON NEURAL NETWORKS, from 2005 to 2010.