Neurocomputing 337 (2019) 354–371

Contents lists available at ScienceDirect

Neurocomputing

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

Deep learning-based methods for person re-identification: A comprehensive review

Di Wu[a], Si-Jia Zheng[a], Xiao-Ping Zhang[a], Chang-An Yuan[b], Fei Cheng[c], Yang Zhao[c], Yong-Jun Lin[c], Zhong-Qiu Zhao[d], Yong-Li Jiang[e], De-Shuang Huang[a,*]

[a] Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Caoan Road 4800, Shanghai 201804, China
[b] Science Computing and Intelligent Information Processing of Guang Xi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning, Guangxi 530001, China
[c] Beijing E-Hualu Info Technology Co., Ltd, Beijing, China
[d] School of Computer Science and Information Engineering, Hefei University of Technology, China
[e] Ningbo Haisvision Intelligence System Co., Ltd., Ningbo, Zhejiang 315000 China

ARTICLE INFO

Article history:
Received 21 March 2018
Revised 8 January 2019
Accepted 9 January 2019
Available online 1 February 2019

Communicated by Prof. Zidong Wang

Keywords:
Person re-identification
Deep learning
Literature review

ABSTRACT

In recent years, person re-identification (ReID) has received much attention since it is a fundamental task in intelligent surveillance systems and has widespread application prospects in numerous fields. Given an image of a pedestrian captured from one camera, the task is to identify this pedestrian from the gallery set captured by other multiple cameras. It is a challenging issue since the appearance of a pedestrian may suffer great changes across different cameras. The task has been greatly boosted by deep learning technology. There are mainly six types of deep learning-based methods designed for this issue, i.e. identification deep model, verification deep model, distance metric-based deep model, part-based deep model, video-based deep model and data augmentation-based deep model. In this paper, we first give a comprehensive review of current six types of deep learning methods. Second, we present the detailed descriptions of existing person ReID datasets. Then, some state-of-the-art performances of methods over recent years on several representative ReID datasets are summarized. Finally, we conclude this paper and discuss the future directions of the person ReID.

1. Introduction

With the development of the monitoring equipment and the increasing demand of public security, quite a few camera networks are installed in public places like theme parks, airports, streets and university campuses. These networks produce huge video image data every day that can be used for forensic purpose or multi-camera tracking. Rely on human monitoring of these video image data may greatly influence the timeliness of multi-camera tracking. For example, in 2012, in the robbery and homicide case of Ke-hua Zhou in Nanjing, China, in order to lock the trajectory of the robber as soon as possible, Nanjing police mobilized more than 500 police to check out tens of thousands of video surveillance images. Facing with such a large number of video images, the police had to continuously work 24 h before finding the robber's tracks. Therefore, the development of the intelligent technology on analyzing the surveillance image is imminent. In computer vision community, intelligent surveillance contains multiple cameras pedestrian tracking, crowd behavior analysis, traffic statistics, activity detection, etc. Person re-identification (ReID) is the most important module of multiple cameras pedestrian tracking, which is defined as the task of identifying whether two pedestrian images captured by disjoint multi-camera views or same camera at different times are the same person or not. The diagram of the task is shown as Fig. 1. Recent years, person ReID has received more and more attentions in computer vision community (Table 1), for the task underpins various crucial applications. A complete person ReID system contains three modules: person detection, person tracking and person retrieval. In this study, we regard the person ReID problem as a task of person retrieval.

Despite the task has extensively studied by researchers worldwide, however, it remains a challenge issue. The pedestrian images captured by non-overlapping cameras usually under an

* Corresponding author.
E-mail addresses: xzhang@ryerson.ca (X.-P. Zhang), dshuang@tongji.edu.cn (D.-S. Huang).

https://doi.org/10.1016/j.neucom.2019.01.079
0925-2312/

Fig. 1. The diagram of person ReID.

Table 1
The number of papers related to person ReID included by the three top conferences in recent years.

	Based On	2010	2011	2012	2013	2014	2015	2016	2017	2018
CVPR	Feature designing	1	0	0	1	2	5	3	3	0
	Distance metric	0	1	1	1	0	1	4	4	2
	Deep learning	0	0	0	0	1	2	5	7	25
ICCV	Feature designing				2		4		3	
	Distance metric				0		2		2	
	Deep learning				0		0		4	
ECCV	Feature designing					2		2		0
	Distance metric					1		3		1
	Deep learning					0		2		18

uncontrolled environment and most of the images are low quality, which leads to some conventional biometrics features like gait and face are not feasible to be used for the task. In these circumstances, the appearance features of the pedestrians, which are extracted from their clothes' colors or objects carried with them, seem to be more suitable for the person ReID task. However, sometimes these appearance features are also helpless. For example, when the colors of individuals' clothes are similar, the color features are invalid to uniquely represent one identity. Moreover, the appearance of a pedestrian can dramatically change under the different camera views when the intensive appearance like viewpoint, pose, lighting, and background clutter changes, which makes the appearances of same pedestrian looked very different. In addition, different pedestrians may share similar appearances, which often leads to the appearances of different pedestrians looked quite similar. In other words, when the camera view changes, the variance of intra-class may be larger than the variance of inter-classes. Fig. 2 shows the matched pairs images from ReID datasets captured by two different cameras, from which we can see that the appearances of the same person vary greatly under the different camera views.

Person ReID is a practical and important task. The existing works handling the task mainly focus on the following aspects:

Traditional methods: (a) designing the hand-craft features that are invariant to poses, illumination, and viewpoints change [1–11]. (b) Learning an effective distance metric to make the distance between the features from different groups farther and the distance between features from the same groups closer [12–19]. (c) Jointly learning feature descriptors and distance metric [20,21].

Deep learning-based methods: using deep learning technology such as convolutional neural network, recurrent neural network and generative adversarial network to address the person ReID. These deep learning-based methods can be roughly divided into six categories, i.e., identification deep model, verification deep model, distance metric-based deep model, part-based deep model, video-based deep model and data augmentation-based deep model. From Table 1, we can observe that almost all

Fig. 2. Sample images from ReID datasets. Each column contains a matched images pair.

methods for person ReID in three top conferences recent years are based on deep learning technology. This motivates us to make a survey of these deep learning-based methods in person ReID community. As to the feature designing-based and distance metric-based methods, we recommend to surveys [22] and [23].

This paper provides an overview of the deep learning based research status for person ReID. Compared with the existing surveys [22–25], we focus on the current deep learning based methods for person ReID in more details. Those methods are currently-prevalent and can reflect future trends in person ReID community. The survey is organized as follows. In Section 2, we review the current deep learning methods designed for person ReID, including identification deep model, verification deep model, distance metric-based deep model, part-based deep model, video-based deep model and data augmentation-based deep model. Section 3 presents the detailed descriptions of existing person ReID datasets. In Section 4, state-of-the-art performances on several representative ReID datasets in recent every year are summarized. Finally, Section 5 concludes the paper.

2. Deep learning for person re-identification

From the computer vision perspective, the ReID can be considered as a matching or multi-classification task [26], let $\delta = \{\delta_1, \delta_2, \ldots, \delta_M\}$ represents a gallery set of M descriptors, then giving a probe descriptor U, the identity of probe person can be formulated as:

$$D = \arg\min_{\delta_i} dis(\delta_i, U),\ \delta_i \in \delta \tag{1}$$

where D represents the identity of U and dis is a suitable distance metric.

Recently, deep learning technology has been widely applied in many computer vision communities [27–31], including face recognition [32,33], object recognition [34,35], object detection [36–38] etc. The technology is a powerful tool that can handle the computer vision tasks without any hand-craft features by an end-to-end way [39–53,53]. Enlightened by the success of this technology, there are a large number of literatures employing deep learning to address the person ReID task. In this section, we give a brief introduction and summary to these methods.

2.1. Identification model

Similar to other multi-class deep learning-based recognition methods, these identification deep models regard the person ReID task as a classification issue, which outputs the corresponding labels of the input person images' attributes. The basic deep architecture of the identification model is shown as Fig. 3. In order to complement the convolutional neural network features, Wu et al. [54] propose a fusion feature network (FFN), which combines a variety of hand-crafted features (e.g. color histogram features and texture features) and CNN features. In the backpropagation phase, the CNN features are constrained by the variety hand-crafted features. The overall network is trained by softmax loss. For reducing the variations of the intra-personal while increasing the differences of inter-personal, in [55], a hybrid deep architecture is proposed for person ReID in which the low-level descriptors including color histograms and SIFT are integrated into the Fisher vectors, and then the Fisher vectors and a deep neural network are combined to produce the finally non-linear features to represent pedestrian images. With multiple domains person ReID datasets, Xiao et al. [56] use a CNN to learn generic deep feature descriptors that suitable for all of these datasets. Specifically, they combine the multiple datasets together to train the designed CNN and propose a domain guide dropout strategy to discard worthless neurons for each domain dataset to keep the deep model in the right track. Experiment results demonstrate that learning deep feature descriptors through using the combined multiple datasets

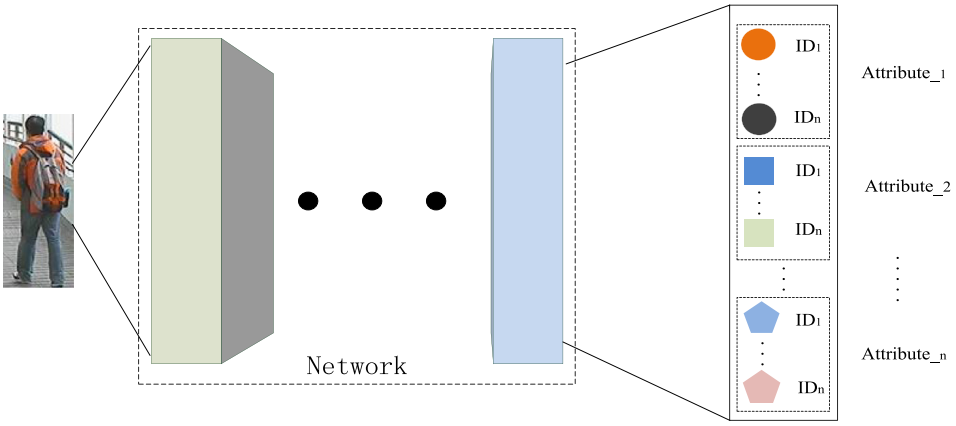

Fig. 3. Identification model.

can improve the performance of the deep model. Recently, Center loss [57] is first proposed to reduce the intra-class variance for face recognition task. The loss maintains a center point for each class and pushes each image embedding to its corresponding center so that the variation between the same image embedding is smaller. To reduce the variation of the intra-class, Jin et al. [58] combine center loss with identification loss to jointly train the designed network. Besides, they introduce a feature reweighting (FRW) layer to learn the weight coefficients of each dimension of the deep descriptors. Similar to the attention mechanism, the FRW layer can pay attention to the useful feature information. However, it seems that the FRW layer easy to lead over-fitting. Zheng et al. [59] propose a pose-box fusion CNN model, which takes pose-box, pose estimation confidence and raw image as input, and then use the fully connected layer to generated the pose invariant embedding (PIE) descriptor to address the pedestrian misalignment problem. Soon afterwards, Zheng et al. [60] propose that pedestrian alignment can be learned by an identification task. Specifically, they design a pedestrian alignment network (PAN) to simultaneously learn the person descriptors and align the person within a bounding box. Lin et al. [61] hold that the person ID recognition learns global representations while attribute classification extracts local aspects, i.e., age, gender, bag and so on. Considering the difference and similarity of the two tasks, they propose combining the attribute classification and identification losses to focus on the local and global aspects of a pedestrian image at the same time. In [62], the authors adopt language descriptions as auxiliary information to promote the visual features learning of re-identification. Specifically, they propose to use local mage-language association strategy to learn semantic consistencies between noun phrases and local visual features, and global image-language association scheme to learn global visual features. There also exist some literatures that simultaneously address the pedestrian detection and re-identification tasks by the deep system. Zheng et al. [63] design a cascaded fine-tuning strategy to train the deep network. Precisely, they first use the strategy to train the detection model and then train the identification model. Besides, a confidence weighted similarity metric that integrates the similarity measurement to the detection scores is applied. To close the gap between real-world scenarios and person re-identification datasets, Xiao et al. [64] design a deep network architecture to jointly handle the person detection and re-identification tasks by an end-to-end way. They introduce a random sampling softmax loss to train the deep architecture under the guidance of unbalance and spare labels. Soon after, Xiao et al. [65] further study the jointly learning system and design an Online Instance Matching (OIM) loss function instead of the usually adopted Softmax loss to train the network. Compared with Softmax loss, OIM is non-parametric. It can exploit the unlabeled pedestrian with a circular queue. But the drawback of the OIM loss is that it trends to easily over-fitting due to its non-parametric. Overall, the input of the identification model is easy to perform and the model can fully use the label information of datasets [23], which is conductive to improve the training efficiency. However, the training object of identification model is inconsistent with its test manner, thus affecting the accuracy of test result. Besides, the scales of current person ReID datasets are quite small, which may lead to the identification model cannot be entirely trained.

2.2. Verification model

Verification model takes a pair of images as input and outputs a similarity value to determine whether the paired images are the same pedestrian or not. The basic deep architecture of the verification model is shown as Fig. 4. Generally, verification model treats person ReID as a binary-class classification problem [66–68]. Li et al. [66] first introduce the verification model to address the person ReID problem. They propose a filter pairing neural network (FPNN) which includes max-out pooling layers and patch-matching to jointly handle geometric transforms and photometric, misalignment, background clutter and occlusions. In the same year, Yi et al. [69] design a “Siamese” deep network for metric learning. The architecture includes three shared parameters independent convolutional networks that perform on three non-overlapping parts of the two images. The deep descriptors of the two input images are produced by the fully connected layers, and then the distance of two output descriptors are calculated by the cosine function. Finally, the cosine function outputs the similarity score. Based on two works above, Wu et al. [68] proposed a PersonNet model for person ReID. The model utilized the patch matching layer proposed by Hirzer [152] to capture local relationship between patches. Moreover, the deep architecture with smaller convolution filters, which is conducive to increase the depth of the architecture. Perhaps the drawback of these verification models above is the depths of their networks are relatively shallow, which is not benefit for digging the deep features that with the ability of discrimination. Besides, the verification model needs to construct images pairs as input and only uses weak datasets label [23], thus reducing the training efficiency. It is worth mentioning that the combination of the identification model and verification model has achieved promising results on person ReID. Chen et al. [70] first propose using the verification and identification losses to train the network for face recognition. To absorb the advantages of the identification and verification model, Zheng et al. [71] propose a verification-identification model that combine the verification

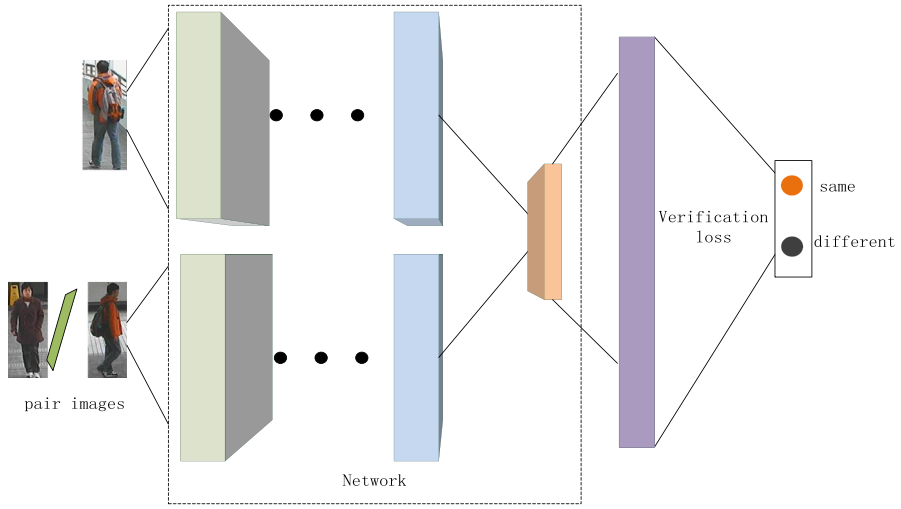

Fig. 4. Verification model.

and identification losses to train the Caffe-Net for person ReID. The literature presents a comprehensive view of the advantages and disadvantages of the two models. Different from the network of face recognition, they use the cross-entropy loss instead of contrastive loss before which the dropout regularization can be adopted on the embedding. Meanwhile, Geng et al. [72] adopt the hybrid strategy and develop a two-steeped fine-tuning approach to transfer the knowledge from large person ReID datasets to the small datasets. Besides, an unsupervised deep transfer learning model based on co-training is proposed to implement ReID without labeled data. To exploit the multi-scale discriminative deep feature, Qian et al. [73] propose a multi-scale deep architecture for ReID. Likewise, the architecture contains two subnets: verification subnet and classification subnets.

2.3. Distance metric-based deep model

Distance metric-based deep model aims to make the distances between the same person images as small as possible while make the distances between different person images as large as possible. The most frequently used metric approach is triplet-based deep architecture. The basic deep architecture of the triplet model is shown as Fig. 5. Triplet model is first proposed by Wang et al. [74] to target the image retrieval task, and then Schroff et al. [75] introduce this triplet model into the face recognition community. Ding et al. [76] first adopt triplet model to address the person ReID task. The input of triplet deep architecture is a triplet unit $I_i = <I_i^1, I_i^2, I_i^3>$, in which I_i^1 and I_i^3 are a mismatched pair while I_i^1 and I_i^2 are a matched pair. For each triplet unit I_i, the model tries to make the corresponding features generated by the deep model satisfies the following condition:

$$\left\|F\left(I_i^1\right) - F\left(I_i^2\right)\right\|^2 < \left\|F\left(I_i^1\right) - F\left(I_i^3\right)\right\|^2 \tag{2}$$

where the $F(I)$ denotes the deep features generated by the network and $\|\|^2$ is the $L2$ norm. Based on the limitation above, they define a triplet loss function for the deep model. The function can be formulated as follows:

$$L(I) = \sum_{i=1}^{n} \max\left\{\left\|F\left(I_i^1\right) - F\left(I_i^2\right)\right\|^2 - \left\|F\left(I_i^1\right) - F\left(I_i^3\right)\right\|^2, C\right\} \tag{3}$$

in which C is a negative constant. Guided by the triplet loss, the network is forced to maximize the distance between the mismatched pair and matched pair under the $L2$ norm (6).

Soon afterwards, Cheng et al. [77] propose an improved triplet loss function for person ReID. They think that Eq. (5) does not specifies how close the deep features generated from same person images should be, thus leading to the instance ascribe to the same person may have a large intra-class distance. Based on this observation, they add a term to the traditional triplet loss, which is:

$$d\left(I_i^1, I_i^2\right) = \left\|F\left(I_i^1\right) - F\left(I_i^2\right)\right\|^2 \leq \partial \tag{4}$$

where ∂ is a margin constant, and that it be much smaller than $|C|$. Under the constraint of this term, the improved triplet loss function can push the different person images farther from each other, and simultaneously pull the same person images closer under the learned deep feature space.

Hermans et al. [78] propose a batch hard triplet loss that adopt a hard mining strategy within a mini-batch. Specifically, they randomly selecting P person identities, and then sampling K images from each P class, thus a mini-batch contains PK images. Finally, they select the hardest negative and the hardest positive samples from one mini-batch to from the triplet units. A semi-supervised attribute learning model is proposed by Su et al. [79], in which the model can learn the mid-level attributes of the person. Concretely, they first train the designed CNN on the dataset with independent attributes labels, and then use the defined triplet loss to fine-tune the CNN on another dataset that only with person IDs labels. Finally, they use the updated CNN to predict the attribute labels of the target dataset. Similar to verification model, triplet-based deep model also has the limitations described below: (1) the model only uses weak label information of ReID datasets [23], (2) the model requires to construct triplet units as input, which reduces the efficiency of training phase. Some approaches optimize the deep architecture by combining the triplet and identification losses [80–82] or combining the triplet and verification losses [83,84]. Liu et al. [80] employ the combination of identification and triplet losses to optimize the designed CNN-LSTM network. The proposed deep model [81] also adopts this combination scheme. Different from [80], in additional to the global identification subnetwork, they add a part-based identification branch to get the part-based representations. Wang et al. [84] analyzes the advantages and limitations of single-image representation (SIR) and the classification of cross-image representation (CIR) in person community, respectively. They propose a deep neural network to simultaneously learn the CIR and SIR. Chen et al. [83] design a multi-task deep network that integrates the triplet and verification losses to take the advantage of the two losses' complementarity for

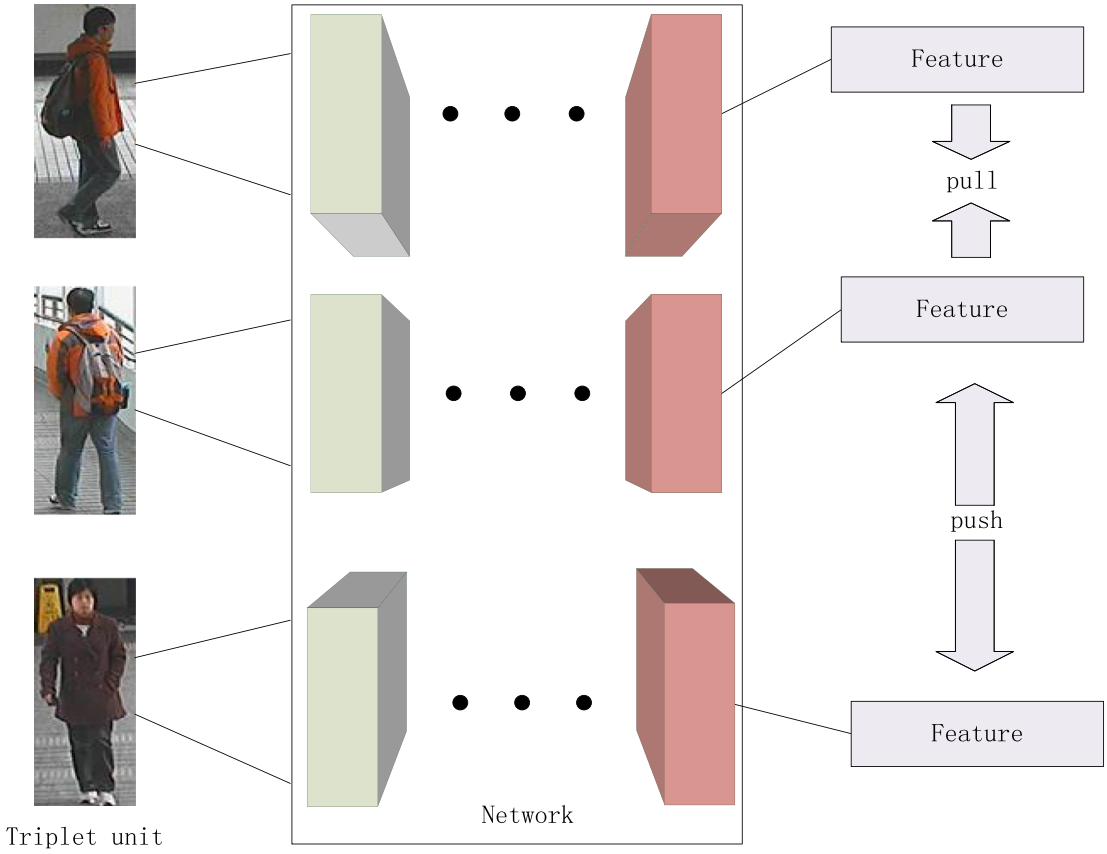

Fig. 5. Triplet model.

person ReID. Moreover, a cross-domain deep model is introduced to improve the performance on small target datasets. Apart from triplet loss, contrastive loss [85], quadruplet loss [86] and margin sample mining loss [87] are also introduced as the loss functions of distance metric-based deep models. Varior et al. [85] design a Siamese CNN with a learnable Matching Gate function, which can diverge the behavior of network during the training and test phases. The architecture is optimized by the contrastive loss, which is:

$$L = Z \times dist_{I_1,I_2}^2 + (1-Z)(\partial - dist_{I_1,I_2})_+^2 \tag{5}$$

where ∂ is a threshold parameter. $(a)_+$ represents max(a,0). I_1 and I_2 are the paired images. Z is the label of the paired images. When the two images are from same person, then $Z=1$, otherwise, $Z=0$. *dist* is the Euclidean distance between the two images.

Chen et al. [86] find that triplet loss-based deep models pay main attentions to get correct orders on the training set, thus suffering from a weak generalization ability on the test set. They introduce a quadruplet loss to address this issue. Compared to the triplet loss, the quadruplet loss can make the output of model with smaller intra-class variation and larger inter-class variation. The loss function can be written as below:

$$L_{quad} = \sum_{i,j,k}^{N} \left[d\left(x_i, x_j\right)^2 - d(x_i, x_k)^2 + \alpha_1 \right]_+ + \sum_{i,j,k,l}^{N} \left[d\left(x_i, x_j\right)^2 - d(x_l, x_k)^2 + \alpha_2 \right]_+ \tag{6}$$

$$s_i = s_j, s_l \neq s_k, s_i \neq s_l, s_i \neq s_k$$

in which α_1 and α_2 are the margin values and s_i is the ID of the image x_i.

Xiao et al. [87] propose a margin sample mining loss (MSML) for person ReID. The loss first calculates a distance matrix, and then select the minimum distance of mismatch pairs and the maximum distance of match pairs to compute the final loss value, thus MSML can utilize the most similar mismatch pair and most dissimilar match pair to train the deep model. In [88], a Hard-Aware Point-to-Set loss is designed to address the sampling issue. Different from the previous schemes, the loss focus on adaptively assigning higher weights to the harder samples.

Overall, distance metric-based deep model can dig the correlation between different person images and learn a similarity measurement in the training phase. Hence, the training target is consistent with its test manner. However, the model needs to construct triple or quadruple units as input and repeatedly computes all possible triplets to select the useful units for network, which reduces the training efficiency. In addition, the model uses weak label information of ReID datasets [23].

2.4. Part-based deep model

There are some literatures [77,89–96] adopt part-based strategy to learn discriminative deep features for ReID. Cheng et al. [77] use a global convolution layer to get the global convolution features, and then they divide the features into four equal individual branches to obtain the part-based deep features. Finally, they concatenate the global and part-based feature vectors to produce the final deep features. Similar to Cheng et al. [77], Li et al. [92] also adopt the division strategy. But they use multi-classification losses to optimize the designed network. In [89], the input images are first resized to 128 × 64 pixels. Then, the images are split into three overlapping parts and each of them with 64 × 64 pixels. Next, they use the three overlapping parts to

input the three individual branches and use a fully connection layer to conclude the deep features of three branches. At last, the output deep feature vector is calculated by another fully connected layer. Ustinova et al. [90] split the raw into three non-overlapping parts and use the multi-region bilinear sub-network for each part. A Siamese LSTM model is proposed by Varior et al. [91] person re-identification. The authors initially divide the image into several rigid parts and extract hand-craft feature like local maximal occurrence and SILTP for each part. Moreover, a LSTM model is introduced to leverage the contextual information of the local descriptors. Li et al. [93] design a *Multi-scale Context-Aware Network* to exploit the global body and local body parts feature. Concretely, instead of using predefined divisiory parts, they propose using spatial transform networks to localize latent person parts. Finally, they concatenate the body parts and full body to form the final representation. Sun et al. [97] propose part-based convolutional baseline for person ReID. The baseline performs unified partition at the convolutional layer to learn part-based features.

Attention mechanism as a part-feature learning module is adopted by Liu et al. [80,99]], Li et al. [98], and Wang et al. [100] to enhance the discriminative ability of the learned deep feature. Liu et al. [80] first introduce the attention model to address the person ReID. They propose a LSTM-based attention model that can dynamically produce part attention feature by a recurrent way for localizing the discriminative local regions of the person image. A attention deep neural network named *HP-net* is proposed by Liu et al. [99], in which a multi-directional attention mechanism is used for capturing multiple attention information from semantic-level to low-level. To handle the misalignment issue in person ReID, Zhao et al. [101] introduce a CNN-based attention model, which utilizes the similarity information of a paired person images to learn the part body for matching. In [98], a Harmonious Attention Convolutional Neural Network is proposed to simultaneously learn feature representations and person ReID selection in an end-to-end way. Specifically, they combine the hard attention and soft attention to learn the region-level and pixel-level parts of the person image, respectively. Wang et al. [100] propose to use 1×1 convolution operation to get an attention mask that can regain spatial structure information for feature maps.

It is noteworthy that the most recent state-of-the-art work on person ReID is based on the part-based deep model. Local visual cues are close to the vision habit of human being and complementary to the global information. The combination of local and global feature is a good choice for person ReID. However, the part-based deep method existing several limitations as follows: (1) adding the part-based branches to the deep model may increase the complexity of model, thus reducing the training efficiency. (2) As mentioned by Li et al. [98], most of attention-based deep models only consider region-level attention and ignore the pixel-level saliency. When facing the small labeled datasets for training and noisy pedestrian images with background clutter and misalignment, these methods are ineffective. (3) As we all know, spatial contextual information is an important element for the discriminative representations, however, only few methods above consider the spatial context information between different part-based features.

2.5. Video-based deep model

Deep learning technology also has been applied to video-based person ReID. Zheng et al. [102] employ two motion features i.e. HOG3D and Gait Energy Image (GEI) as well as CNNs to learn the discriminative embedding under the pedestrian subspace. They find that motion features are less effective when facing the large datasets with occlusion, complex background and various pose. Instead, deep feature has good performance on the video datasets with large amount of training data. Mclaughlin et al. [103] first utilize a CNN model to extract features from the video sequence, and then they employ a Recurrent Neural Network (RNN) to transform information between various frames. Wu et al. [104] design a recurrent DNN model for video-based person ReID. They first use a CNN incorporating a recurrent layer to extract features. Then, they use temporal pooling to combine all time-steps features to get a global appearance feature for the entire sequence. Yan et al. [105] propose a sequential/progressive fusion model in which the hand-craft features like LBP and color are fed into the LSTM network. Liu et al. [106] hold on that the single-stream CNN model cannot make use of the valuable of temporal information and only use appearance features. Based on this point, they use two individual streams to process the temporal and spatial information, respectively. Then they use the designed certain intermediate to fuse the two kinds of information. In addition, an Accumulative Motion Context (AMOC) network is introduced to capture the motion context and spatial features from the video sequences of pedestrians. Li et al. [107] propose a deep feature guided pooling (DFGP) model. They use a PCA-based CNN to generate the deep features, and then they utilize a max pooling to aggregate the hand-craft features to enhance the motion variations of different pedestrians. Finally, these deep features and hand-craft features are combined to compose the final representations. In [108], a two stream CNN is proposed to simultaneously capture temporal and spatial information. Xu et al. [109] consider the interaction between the paired sequences. They introduce an attention spatial-temporal pooling network in which the temporal attention weights of one sequence is guided by the features of another sequence through a sharing matrix. Chen and co-workers [110] introduce co-attentive snippet embedding and similarity aggregation for the video-based person ReID. They divide the long sequences into multiple short fragment and aggregate the top-ranked fragment for estimating the similarity of the sequences. In [111], a deep association learning method is proposed to address the unsupervised video-based ReID. Through optimizing two margin-based association losses, the model limits the association of per frame to the best-matched cross-camera representation and intra-camera representation.

2.6. Data augmentation-based deep model

At present, the number of images for one person in a certain dataset is still limited in person ReID community. For example, the average numbers of per person for the large-scale ReID datasets like CUHK03, Market-1501 and DukeMTMC-reID are 9.6, 17.2 and 23.5, respectively. Using such scale datasets to train the deep model may lead to over-fitting issue. Therefore, some literatures [112–115] attempt to extend the training set of the ReID datasets. Zheng et al. [112] first introduce a generative adversarial network (GAN) to generate unlabeled pedestrian samples and adopt a CNN sub-model for feature representation learning. Since the generated pedestrian images has no labels, through using a label smoothing regularization for outliers (LSRO) method, the model mixes the unlabeled GAN data with the real labeled data for training. For person ReID dataset, the image style of different cameras in it may different, which leads to the pedestrian images captured by non-overlapping cameras suffer from intensive changes in background and appearance. In order to address this problem, Zhong et al. [113] introduce a camera style adaption model to adjust the CNN training. More specifically, they use the CycleGAN [116] to transfer the style of images captured by one camera to another. Given a training sample from one certain camera, the model can produce new images with the style of other cameras. Besides, to alleviate the noise of generated image caused by CycleGAN, they introduce label smooth regularization to the new images. So far, different person ReID datasets exist domain gap. The gap would result in the serious performance drop when model training on

certain dataset and testing on another dataset. Aiming at this problem, Wei et al. [114] propose a person transfer GAN model, which makes up the domain gap via transferring pedestrians in *C* dataset to *D* dataset. After transferring the pedestrians from *C*, the transferred images keep their *IDs* and hold the similar styles like lightings, backgrounds, etc., with dataset *D*. In [115], the authors introduce a pose-normalization GAN model to alleviate the influence of pose variation. Given a pedestrian image, the model uses a desirable pose to generate a composited image of the same *ID* with the initial pose replaced with the desirable pose. Then they use the pose-normalized images and original images to train the ReID model to produce two sets of features, respectively. Finally, the two types of features are fused to form the final descriptor. Conclusively, GAN-based data augmentation method could enhance the generalization capacity of ReID model and solve the difficult of person ReID from a certain standpoint to some extent. But the quality of generated images is relatively poor, thus bringing noise to the ReID system. Future work may focus on designing the GAN model that can generate high quality images.

2.7. Other perspectives

Apart from the six major types of deep learning-based models, there have been some deep learning-based researches on person ReID from other perspectives, such as camera network based methods, open-set person ReID, semi-supervised learning-based person ReID, low-resolution person re-identification and so on. In this section, we give a brief introduction of these types of methods.

In ReID community, the model trained on one specific dataset is directly tested on another one, the performance of it dramatically drops because the dataset bias. This issue is called domain adaption ReID. Some deep models are also proposed to address the domain adaption person ReID issue [117–122]. To address the open world nature and dynamic of the ReID issue, Panda et al. [117] design an unsupervised adaptation strategy for person ReID under a dynamic camera network, in which the existing system may insert a new camera to get extra information. In order to learn the deep features from the ReID dataset with no or a few labels, Fan et al. [118] introduce a progressive unsupervised learning model to transfer the pre-trained deep architecture to the unlabeled target domains. In [119], an unsupervised domain adaption deep model named Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) is proposed to learn feature descriptor through transferring the attribute knowledge and labeled information of source domain to the unlabeled (unseen) target domain in an unsupervised way. Deng et al. [120] propose a "learning via translation" architecture to translate the labeled pedestrian images from source domain to the unlabeled target domain. Moreover, they introduce unsupervised domain-dissimilarity and self-similarity for similarity preserving image generation to preserve the latent identity information for the foreground person. Zhong et al. [122] also focus on domain adaptation for person ReID. They propose a Hetero-Homogeneous Learning algorithm which jointly considers domain connectedness and camera invariance to learn more generalization pedestrian embedding for the unlabeled target domain. Both the person ReID methods mentioned above focus on camera pairwise re-identification. However, these methods rarely consider of maintaining consistency of the results across the camera networks, which leads to inconsistent associations when the results from different paired camera are integrated. To address this problem, Lin et al. [123] present a consistent-aware deep learning model to obtain the maximal correct matches for person ReID under a camera network. They integrate the consistent-aware information to a designed deep learning architecture, in which the deep features and image matching are learned jointly. Besides, a gradient descent-based algorithm is presented to get the globally optimal matching. The generic person ReID issue basically presumes that the gallery set contains the target probe image. However, this hypothesis is always not true in practical scenarios. Because the probe population may extremely huge and the probe set may include mainly non-target people. In this situation, it is not practical that the probe people is supposed to be one target person. Such case is called open-set person ReID. Deep technology used for open-set person ReID is rarely. Most recently, Li et al. [124] propose a deep open-set group-based person ReID approach, which adopts the adversarial learning strategy to relieve the attack of similar non-target person. Their work uses target-like samples generated by GAN to attack the feature extractor, and makes the extractor can tolerate the attack through discriminative learning. Through the adversarial strategy, person ReID system is more stabilized when facing the open-world issue. To alleviate the label-dependence of supervised methods, there have some studies [125–127] focus on semi-supervised person re-identification. Zheng et al. [126] use the DCGAN to generate unlabeled sample to augment the scale of datasets. They assign a disributed pseudo-label (MpRL)to the generated samples for semi-supervised learing. In [127], the authors further propose a multi-pseudo regularized label method to label the generated samples, which shows the possibility of the subordination relation between the generated images and all predefined training classes. Ding et al. [125] emphasize that the labeling method for the GAN samples should take into consideration the representation similarity and the relationships between their and real samples. They introduce a approach called FAPL with distributed and one-hot label encodings for semi-supervised ReID learning.

In open world person ReID, the captured pedestrian images usually have low resolutions (LR), which causes to the resolution mismatch dilemma when matching the high resolution images in gallery set. To address this dilemma, several studies [128,129] focus on solving the low-resolution person ReID. Wang et al. [128] further design a cascaded super-resolution GAN architecture to address the SALR-ReID issue. Specifically, they cascade multiple super-resolution GAN modules in series to joint handle the scale-adaptive upscaling and super-resolving LR operations. Besides, they integrate a ReID network to promote the image feature representation learning. In [129], a combined deep CNN model is proposed for improving the performance of cross-resolution ReID. Apart from the accuracy of re-identification, the searching speed of the person ReID is also important in practical application. Zhu et al. [130] and Wu et al. [131] integrate hashing and deep learning into an uniform architecture to joint consider the accuracy and efficiency for large-scale person ReID.

The most frequently used sensor is RGB in person ReID domain. However, RGB sensor is sensible to lighting, occlusions and clutter conditions. When these conditions change, the RGB appearance-based models toward to misfire. To overcome this issue, a few depth-based [132] models and cross-modality-based [133–137] models have been developed in recent years. Munaro et al. [138] exploit skeleton-based feature based on the skeletal tracking method to warp pedestrian point clouds to a standard pose, and then fuse a set of warped point clouds from different frames to compose the model. Haque et al. [132] introduce a recurrent attention model for depth video person ReID problem, in which 4D spatio-temporal signatures are learned from a high-dimensionality 4D input space. Using RGB-D sensors, Pala et al. [133] merge anthropometric measures taken from depth data to clothing appearance descriptors to improve the accuracy of descriptors. They introduce a dissimilarity-based architecture to fuse and build the multi-modal descriptors of person samples for ReID mission. A multiple modality ReID dataset called *SYSU-MM01* is created by Wu et al. [136] for cross-modality problem. Besides, they propose deep zero-padding to train one-stream framework

Table 2
Accuracy of several loss functions with two base models.

Models	Losses	Market-1501			CUHK03			Duke-MTMC		
		mAP	Rank=1	Rank=5	Rank=1	Rank=5	Rank=10	mAP	Rank=1	Rank=5
Inception	Softmax	51.2	75.5	88.9	72.9	91.7	95.6	34.3	54.7	63.1
	OIM	53.9	77.6	90.6	77.9	93.3	96.1	41.2	60.9	68.4
	Triplet	61.7	79.4	91.2	80.3	94.8	96.4	48.3	65.3	72.3
	MSML	61.9	80.1.	91.1	82.2	95.9	96.6	49.1	64.9	73.1
ResNet-50	Softmax	60.0	80.9	92.5	71.2	90.5	95.2	41.2	63.1	70.3
	OIM	61.1	82.3	93.3	77.9	94.1	96.3	47.8	70.3	73.2
	Triplet	67.6	84.8	94.1	84.1	95.3	97.1	53.9	73.6	80.4
	MSML	69.6	85.2	93.7	84.0	96.7	98.2	53.3	74.5	81.8

for cross-modality matching. In [134], Dai et al. propose a cross-modality GANs model to investigate the ReID between RGB and infrared images. The GANs use CNN to learn image representations and use a modality classifier to discriminate between infrared and RGB image modalities. Hafner et al. [137] introduce a cross-modality distillation network for person ReID between RGB and depth images, in which the learned representations can be transfer from one sensor to another.

3. Results for loss functions with base models

We use the codes provided by Open-ReID repository to implement comparisons experiments to estimate the performance of several loss functions with two base models, i.e., Inception and ResNet-50. We conduct the experiments on three large-scale datasets and the results are shown in Table 2. Since the number of images for per pedestrian is limit in the existing person ReID datasets, thus using the identification model with Softmax loss can be easy to cause over-fitting. From Table 2, we can observe that the performance of Softmax loss is relatively poor among the four losses. For OIM, it is an improved version of identification loss, compared to Softmax loss, the mAP and Rank-1 accuracies are all increase on three datasets with the two base models. Both triplet and MSML losses have better performance than identification losses (i.e., Softmax and OIM). MSML is a little better than triplet on Market-1501 while triplet is better on CUHK03, and the two types of losses have an equally-matched performance on Duke-MTMC. Compared to identification model, these distance metric-based deep model can learn a distance metric for the image pairs, thus the training objective of the model is consistent with its testing manner. However, the inputs of the distance metric-based model are difficult to organize and the model only use weak label information of the datasets, which reduces the training efficiency. Also, from the discussion above, we find that the identification loss and distance metric-based loss have complementary advantages to some extent, which indicates that jointly adopt these two types of loss functions is a good choice for the CNN architecture.

4. Datasets

To exploit robust person ReID models, it is crucial to have the available ReID datasets with the characteristics of cluttered background, occlusions and overlapped bodies, etc. So far, several available datasets for person ReID have been released. Among them, ViPER [139], CUHK03 [66], PRID 2011 [140] and Market-1501 [141] are the most commonly used benchmark datasets for person ReID evaluations. These datasets can be roughly divided into three categories: RGB image-based datasets (Fig. 6), video sequence-based datasets (Fig. 7) and cross-modality datasets (Fig. 8-11). In this section, we give a brief introduction to these person ReID datasets.

4.1. RGB image-based datasets

ViPER. The ViPER dataset [139] contains 632 identities, and each identity has two images captured from different non-overlapping cameras. The dataset with large variations in background, lighting conditions and viewpoint and it has 28 different viewpoint angle pairs and eight similar viewpoint angles. Hence, it is collected to test viewpoint invariant person ReID model. All images in the dataset are cropped into 128×48 pixels. This dataset is one of the most challenging datasets for automated person ReID.

GRID. The dataset is collected by eight non-overlapping fields-of-views cameras from an underground station [142]. The images of this dataset have large illumination variations and are low resolution. This dataset has 1275 images in which the 250 pairs of person images captured from two different cameras and the rest 775 images are taken by a single camera.

CUHK01. CUHK01[12] consists of 3884 images of 971 pedestrians with two surveillance camera views. One view catches the back or frontal view of a pedestrian, and another captures the pedestrian's profile views.

CUHK03. The dataset is one of the largest person ReID datasets which contains 13,164 images of 1360 identities. All identities are taken from six camera views, and each pedestrian is captured by two cameras. This data set provides two setting. One automatically annotated by a detector and the other manually annotated by human. Among the two setting, the former is closer to practical scenarios.

Market-1501. This dataset consists of 32,643 annotated boxes of 1501 persons. Each pedestrian is collected by at least two cameras and at most six cameras from the front of a supermarket. The boxes of pedestrians are captured by the Deformable Part Model (DPM) detector.

DukeMTMC-reID. The DukeMTMC-reID which is created for image-based person ReID is a subset of DukeMTMC dataset. It consists of 36,411 pedestrian images that belong to 1812 identities taken from eight high-resolution surveillance equipments. Among these 1812 pedestrians, 1404 of them captured by more than two camera views and the rest of them are regarded as distractor identifications.

Airport. The dataset is collected from six cameras of an indoor intelligence surveillance system in an airport. It contains 9651 people of 39,902 bounding box images, with average 3.13 images

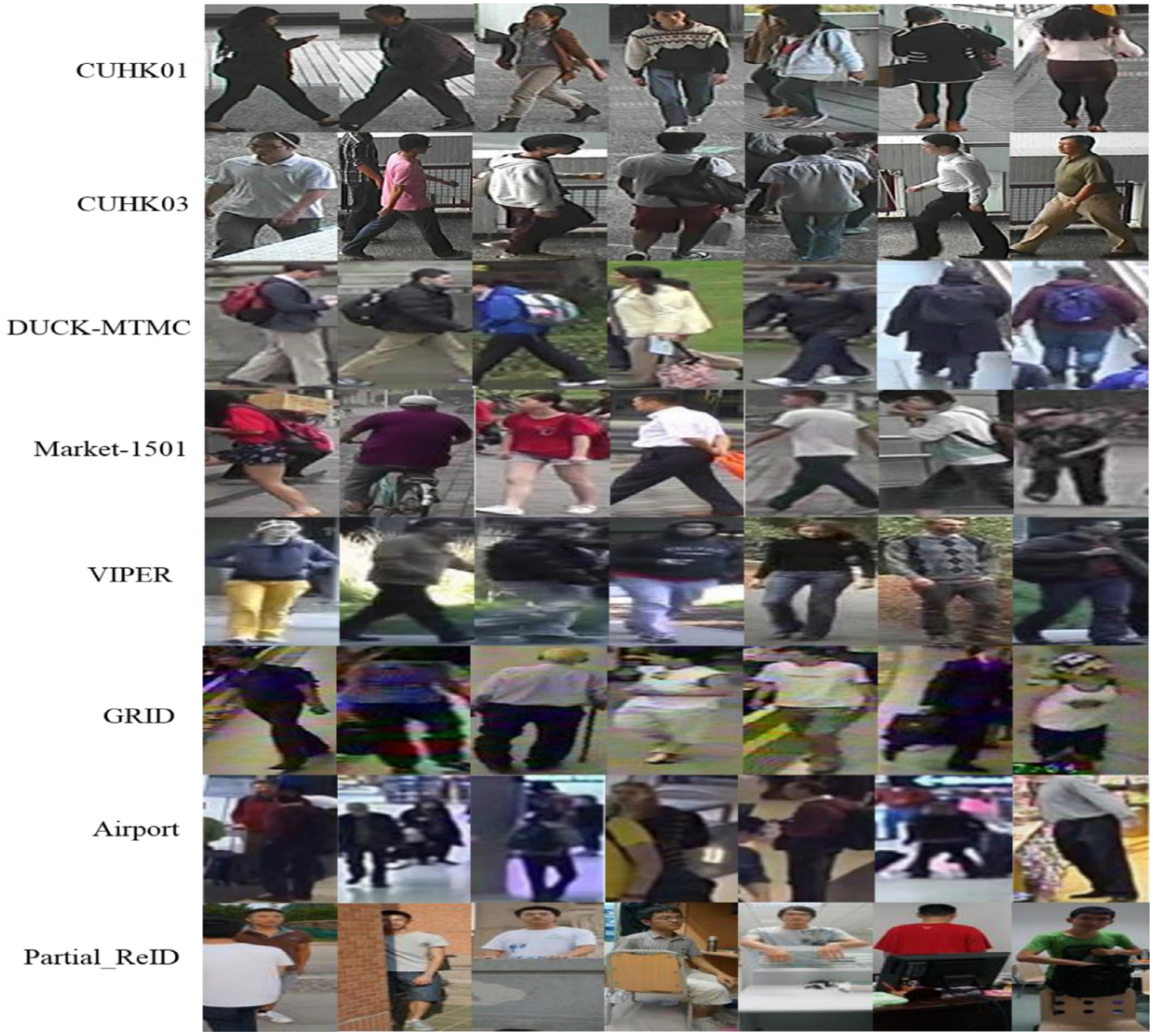

Fig. 6. Sample images from the frequently used seven image datasets.

for per pedestrian. The sizes of the bounding boxes range from 54 × 130 to 166 × 403. Among the 9651 persons, 1382 of them are paired within at least two cameras. More specific details about this dataset can be found in the references [143,144].

Partial-reID. This dataset [145] is specially created for partial person ReID, which contains 600 images of 60 pedestrians, with five partial images and five full-body images for per pedestrian. These images are captured from different backgrounds, viewpoints and types of occlusions.

4.2. Video sequence datasets

All of the datasets mentioned above are used to test image-based person ReID models, there also have some datasets that are used for testing video-based person ReID models, e.g. ETHZ [146], PRID-2011 [140] and iLIDS-VID [147]. Fig. 7 displays some example images of these datasets.

3DPeS. The 3DPeS [148] dataset consists of a set of selected snapshots that contains 1011 images and 192 pedestrians. As shown in Fig. 7, the dataset suffers from illumination and viewpoint variations.

ETHZ. The dataset [146] contains three video sequences taken by two moving cameras at a crowded street. The first sequence consists of 4857 images of 83 identities. The second sequence includes 21,961 images of 35 identities and the third sequence has 1762 images of 28 persons. All images are cropped into 128 × 64 pixels. This dataset involves significant illumination variations and occlusions.

PRID-2011. The images of this dataset [140] are captured from two non-overlapping surveillance cameras. One camera captures 749 pedestrians and the other camera captures 385 pedestrians. Among these pedestrians, 200 persons recorded in both cameras. All images are cropped into 128 × 48 pixels. Different from other datasets, PRID 2011 is captured in a relatively clean and simple scene and the dataset has consistent illumination changes.

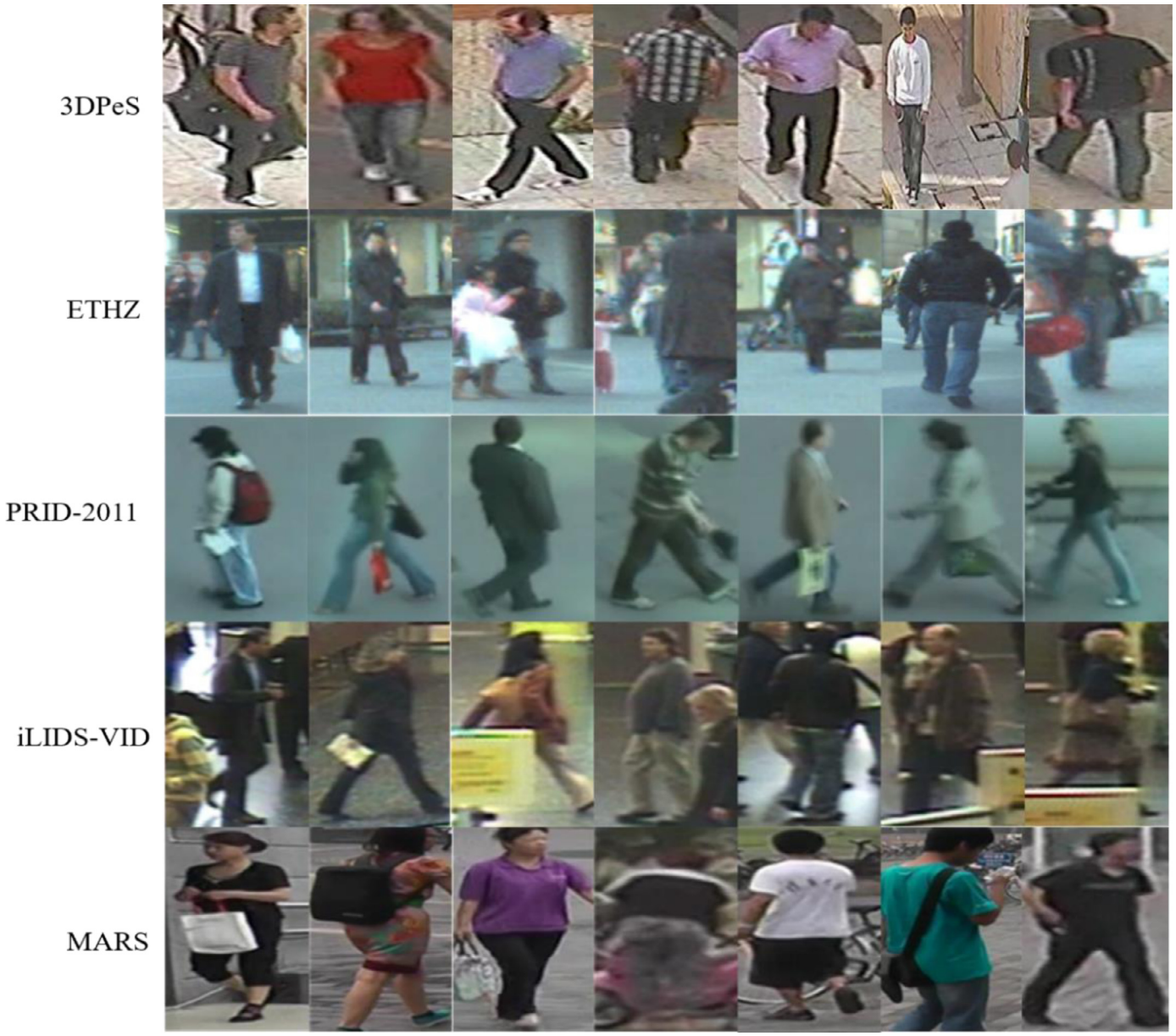

Fig. 7. Sample images from the frequently used four video sequence datasets.

iLIDS-VID. The dataset [147] consists of 600 videos and total 300 persons. For each pedestrian, it has a pair of videos captured from two non-overlapping cameras. The length of each video ranges from 23 to 192. This dataset is collected by a multi-camera CCTV network from an airport arrival hall, which is challenging due to viewpoint and lighting changes, clothing similarities and background occlusions.

MARS. The MARS dataset [149] consists of 1261 identities and about 2000 videos, which make it to be the largest video-based person ReID dataset. The sequence is automatically captured by the GMMCP tracker and DPM detector. These videos are captured by at least two camera views and at most six camera views, and an average of 13.2 video sequence for each individual. In additional, the dataset contains 3248 distractor sequences.

4.3. Multi-modal datasets

RGBD-ID. This dataset [150] consists of four different groups of data. Both the four groups contain the same 79 pedestrians. The first group records the pedestrians from a frontal view, where the pedestrians are away from the camera at least two meters. The second and third groups are recoding the 79 pedestrians normally walking and the back view of the pedestrians walking are recorded by the fourth group. This dataset is created in different days and the visual aspects of the pedestrians may change. Fig. 8 displays some example images of this dataset.

BIWI RGBD-ID. This dataset [151] targets to long-term person re-identification from RGB-D cameras. It consists of 50 training and 56 test sequences for 50 different persons. The dataset includes RGB images, pedestrians' segmentation maps, depth images and skeletal data. Among these 50 persons, most of them wear different clothes in the training and test sequences. Fig. 9 displays some example images of this dataset.

KinectREID. This dataset [133] is created by using the official Microsoft SDK and Kinect V1 sensors. It contains sequences of 71 pedestrians taken from the authors' department. Each pedestrian involves seven video sequences, with the corresponding skeleton points and segmentation maps. The dataset provides three

Fig. 8. Sample images from the RGBD-ID dataset. Note that the visual appearances of pedestrians change since they are captured in different days.

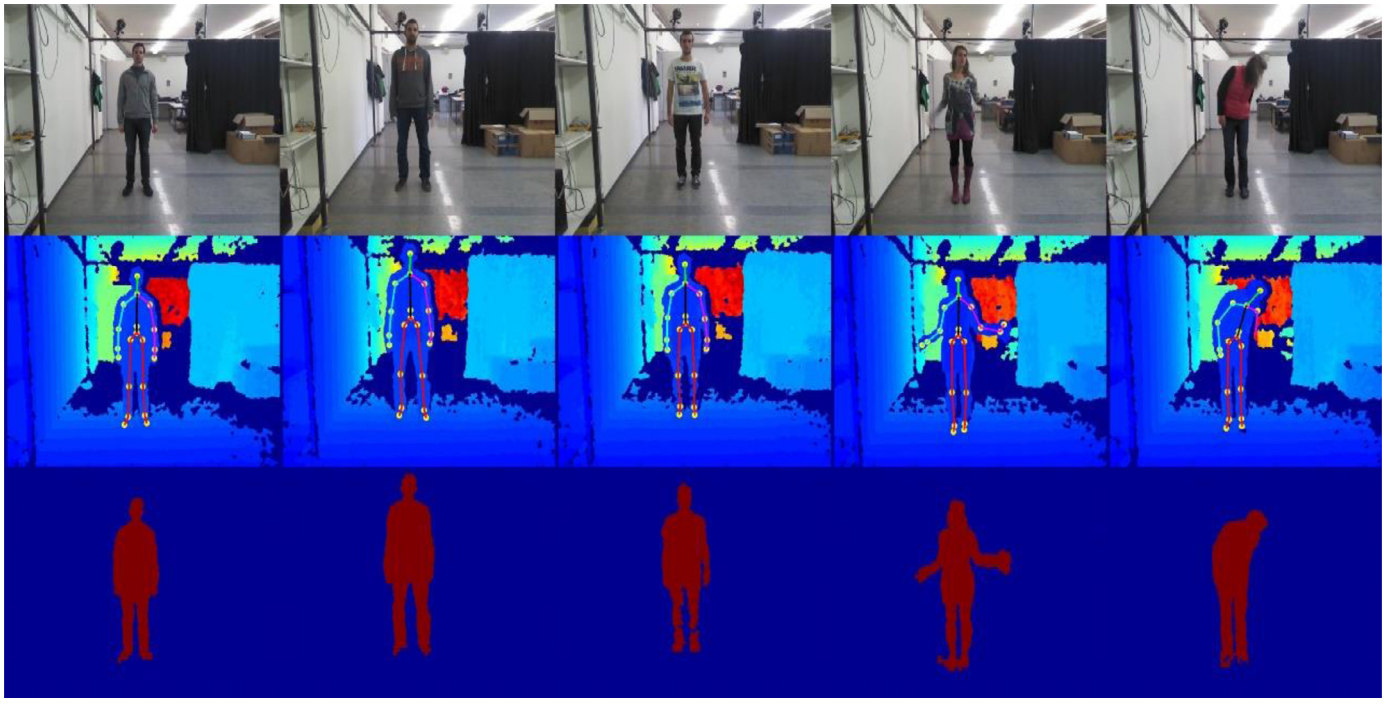

Fig. 9. Sample images from the BIWI RGBD-ID dataset.

view points: near-frontal view, near-rear view and lateral view. Fig. 10 displays some example images of this dataset.

SYSU RGB-IR. The SYSU RGB-IR ReID dataset [134] is captured by six cameras, including four RGB cameras and two infrared cameras. This dataset consists of 491 pedestrians with 15,792 infrared images and 287,628 RGB images. And the training set contains 12,792 infrared images and 19,659 RGB images. This dataset is challenging because the great differences between the infrared and RGB modalities. Fig. 11 displays some example images of this dataset.

Through the descriptions of these datasets above, we can conclude the developing trends of the datasets in person ReID community. (a) The volume of the dataset trends to increase, especially for the RGB image-based datasets. For example, from CUHK03 to Market-1501 to DukeMTMC-reID, both the IDs and the number of instances for per person are increased. (b) The modality of data become more diversity. Some datasets such as RGBD-ID and BIWI RGBD-ID contain both RGB and depth images. (c) The dataset with specific purposely emerging. There have been some datasets like Partial ReID and Partial-iLIDS specifically designed for partial person ReID. Both these changes are helpful for promoting the practical performance of the person ReID models.

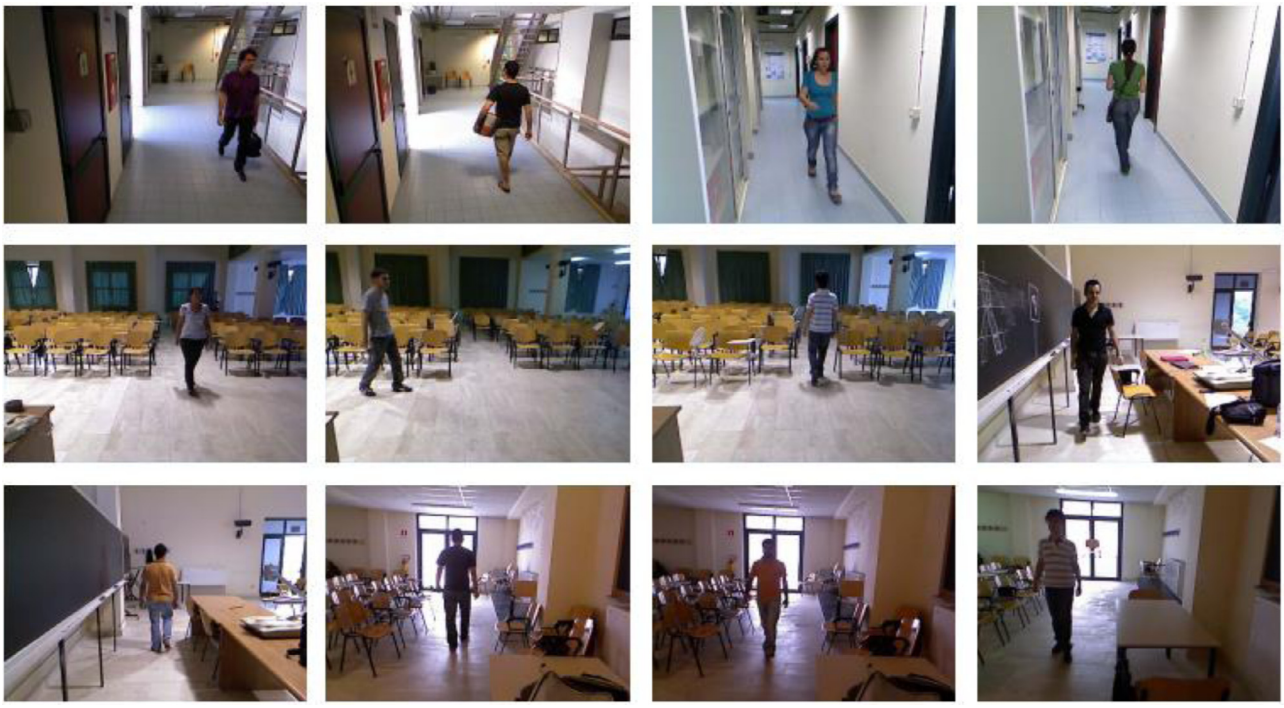

Fig. 10. Sample images from the KinectREID dataset.

Fig. 11. Sample images from the SYSU RGB-IR ReID dataset.

5. State-of-the-art performances over recent years

We give a summary of the ReID accuracy of some state-of-the-art methods on several well-known datasets mentioned above. Table 3 presents state-of-the-art models on these datasets in recent every single year. To have a comparison, we also present the performances of feature designing-based methods and distance metric-based methods on the datasets.

As shown in Table 3, the presented datasets in Table 3 are VIPeR [139], CUHK03 [66], iLIDS-VID [153], PRID 2011 [140], Market-1501 [141], Partial-ReID [145] and SYSU RGB-IR [134]. From Table 3, we can observe that both the accuracies of re-recognition of the seven datasets are increasing during those years. More specifically, from the years 2015 to 2018, the Rank-1 accuracy of the Market-1501 dataset have increased from 14.2% to 93.6% under single-query mode, with an increasing of 79.4%. As to the smallest dataset, ViPER, from the years 2011 to 2018, the performances of state-of-the-art methods have improved from 21.8% to 63.9%, with an increasing of 42.1%. For the PRID 2011 and iLIDS-VID datasets which are used for video-based person ReID models, the Rank-1

Table 3
Accuracy of the state-of-the-art models on popular person ReID datasets over the recent years.

	Based On	2011	2012	2013	2014	2015	2016	2017	2018
ViPER	Feature designing	21.8 [152]		30.0 [153]	37.8 [154]	63.9 [155]	53.5 [156]		
	Distance metric		27.0 [157]						
	Deep learning							53.8 [158]	51.9 [159]
CUHK03	Feature designing			8.8 [160]					
	Distance metric		14.2 [157]						
	Deep learning				20.7 [66]	54.8 [67]	85.4 [72]	88.5 [161]	94.9 [162]
iLIDS-VID	Feature			10.2 [160]	34.5 [147]	44.3 [163]			
	designing								
	Distance metric								
	Deep learning						69.1 [164]	68.7 [106]	85.4 [165]
PRID 2011	Feature designing			25.8 [160]	37.6 [147]	64.1 [163]			
	Distance metric								
	Deep learning						66.8 [164]	83.7 [106]	93.0 [165]
Market-1501	Feature designing					44.4 [141]			
	Distance metric								
	Deep learning						83.7 [72]	84.9 [78]	93.6 [166]
Partial-ReID	Feature designing								
	Distance metric					36.0 [145]			
	Deep learning								43.0 [167]
SYSU RGB-IR	Feature designing								
	Distance metric								
	Deep learning							14.8 [136]	29.1 [134]

accuracy on them have increased from 25.8% to 93.0% and from 10.2% to 85.4%, respectively. On Partial-ReID and SYSU RGB-IR datasets, the accuracy rates increase of 9.0% and 14.3%, respectively. In addition to the Rank-1 accuracy, some deeper clues can be found from Table 3. Since VIPeR datasets are relatively small, the advantages of the deep learning technology cannot be completely used. The most recent state-of-the-art methods on this dataset are based on distance metric approach. Apart from the VIPeR dataset, both the most recent state-of-the-art methods on the other six datasets are based on deep learning technology. On the two largest RGB image-based datasets, i.e., CUHK03 and Market-1501, the deep learning methods are overwhelmingly superior to both the feature extraction methods and distance metric methods. Compared to the RGB image-based ReID datasets, the scale of video-based ReID training data is obviously larger, there is no doubt that the performances of the deep learning methods achieve better performances than hand-craft methods on the datasets. Since occlusion and multi-modality data issues are new topics in person ReID community, we can see that the performances of state-of-the-art models on partial-ReID and SYSU RGB-IR datasets are relatively poor. The best Rank-1 accuracies on partial-ReID and SYSU RGB-IR datasets are 43.0% and 29.1%, both based on deep learning methods. From the accuracy performances on these seven datasets, we can observe there are still have space to improve, especially for the video-based datasets and the new specific purposely datasets. On the RGB-based datasets, the rank-1 accuracy is relatively high. For instance, on the CUHK03 dataset, the Rank-1 accuracy of most advanced model achieves 94.9%, which exceeds the limit of human being's observation. However, these RGB-based datasets exist limitations, which leads to the model trained on them with poor robustness in practical scenarios. To address this issue, new dataset closer to the real scene should be created. Besides, in term of the rank-1 accuracies on these existing RGB-based datasets, we speculate that the state-of-the-art performances are most probably based on deep learning systems over the next few years.

6. Conclusion and future directions

So far, person ReID is still a challenging task since the inherent limitations like low resolution images, illumination variation, viewpoint variant, etc. The task has received great attention over recent years, and quite a few person ReID models have been proposed for it. In this paper, a survey of current methods for person ReID is presented. First, we discuss the meaning and challenging issues of the task. Second, current deep learning-based methods are briefly reviewed. We roughly categorize these deep learning-based methods into six groups: i.e., identification model, verification model, distance metric-based deep model, part-based deep model, video-based deep model and data augmentation-based deep model. Third, we provide the detail descriptions of several popular available ReID datasets. Finally, we summarize the accuracies of some state-of-the-art methods on seven person ReID datasets over recent years.

Based on our survey above, we present several future research directions from personal opinions.

(1) In part-based deep learning method, most of them combine part and global branches to obtain the final descriptors, which affects the efficiency of testing. How to integrate the part and global information into a unitary descriptor and use one branch for testing is a research direction for part-based deep model. Besides, most part-based ReID work do not consider the interdependencies between local features. Future part-based deep model is advised to introduce relational information such as spatial context and temporal information to the local features.
(2) In global deep features learning method, existing methods usually use the person images contain complex background to directly learn feature representation, which introduce irrelative background information for ReID. One of the ideas that can alleviate to this issue is introduce the human body mask learning branch to force the backbone network to automatically locate the human body part for feature learning.
(3) GAN-based deep model served as a data augmentation technology can overcome the limitations of existing person ReID datasets to some extent. But the qualities of most of synthesized person images generated by the current GAN-based models are not high, which brings noise to the original datasets. Future GAN-based data augmentation should consider generating more quality samples for the datasets. Moreover, both the existing GAN-based data augmentation methods are designed for image-based ReID datasets. Another promising direction is to design GAN model to generate sequence sample for the video-based datasets.
(4) Most of the current methods assume that the bounding box of the person is well given. In other word, the person detection is accurately performed before retrieval. However, in the real scene, the bounding box of the pedestrian often with a certain deviation, and the quality of detection may affect the performance of ReID. To address this problem, several literatures propose to handle the detection and re-identification jointly, but these methods rarely try to analyze and solve the effect of detection deviation on ReID performance. Therefore, integrating the detection and ReID tasks into a unitary framework and analyzing the effects on ReID performance caused by accuracy of detection are still open to challenge in the future direction.
(5) In real-world scenarios, the surveillance equipment may be diversity, which leads to the captured images are multimodal. In this situation, RGB image-based methods are invalid. So far, only few studies pay attention to this issue. And the performance of the current methods is still far from satisfaction. Therefore, how to use the deep learning technology to learn discriminative information to search the pedestrian between multimodal image data is another direction for person ReID.
(6) Although the current models have achieved encouraging progress on person ReID datasets, they are narrow to apply to realistic scenarios, for both of them try to solve the short-period ReID and do not considering the robustness among the different cameras over a long period. Furthermore, compared to other computer vision tasks, the scales of current person ReID datasets are quite small. For example, in face recognition community, the CelebA [154] dataset contains 202,599 images of 10,177 identities. Caltech datasets used for pedestrian detection includes 350,000 pedestrian boxes of 2300 identities. However, the commonly used large scale dataset for person ReID only contains about 30,000 images, which constrains to learn generalization models. Besides, the modality of data may be multifarious in the real world, only rely on RGB image train model has limitation. To address these issues, future person ReID dataset should take into account the following factors: (a) long term dataset collected over several days. (b) Larger scale with more identities and more instances for per identity. (c) Introduce multi-modality data to the dataset.

All in all, person ReID is a quickly developing field with challenge and opportunity. We hope this paper can be helpful for new researchers quickly understand the topic.

Acknowledgments

This work was supported by the grants of the National Science Foundation of China, Nos. 61520106006, U1611265, 61532008, 61672203, 61572447 61861146002, 61732012, 61772370, 61702371 and 61672382, China Postdoctoral Science Foundation, Grant No. 2016M601646, and supported by “BAGUI Scholar” Program of Guangxi Province of China.

References

[1] J. Sivic, C.L. Zitnick, R. Szeliski, Finding people in repeated shots of the same scene, in: Proceedings of the 2006 British Machine Vision Conference, Edinburgh, UK, September, 2006, pp. 909–918.
[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2005, pp. 886–893.
[3] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vis. 61 (2003) 55–79.
[4] S. Khamis, C.H. Kuo, V.K. Singh, et al., Joint Learning for Attribute-Consistent Person Re-Identification[C] European Conference on Computer Vision, Springer, Cham, 2014.
[5] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2014, pp. 2528–2535.
[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Proceedings of the Computer Vision and pattern Recognition, 2010, pp. 2360–2367.
[7] N. Martinel, C. Micheloni, G.L. Foresti, Saliency weighted features for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 191–208.
[8] L. An, X. Chen, S. Liu, Y. Lei, S. Yang, Integrating appearance features and soft biometrics for person re-identification, Multimed. Tools Appl. 76 (2017) 1–15.
[9] H.M. Hu, W. Fang, G. Zeng, Z. Hu, B. Li, A person re-identification algorithm based on pyramid color topology feature, Multimed. Tools Appl. (2016) 1–14.
[10] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1622–1634.
[11] B. Ma, Y. Su, F. Jurie, Covariance descriptor based on bio-inspired features for person re-identification and face verification ☆, Image Vis. Comput. 32 (2014) 379–390.
[12] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: Proceedings of the Asian Conference on Computer Vision, 2012, pp. 31–44.
[13] P.F. Felzenszwalb, D.A. Mcallester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: Proceedings of the Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[14] L.D. Bourdev, J. Malik, Poselets: body part detectors trained using 3D human pose annotations, in: Proceedings of the International Conference on Computer Vision, 2009, pp. 1365–1372.
[15] M. Andriluka, S. Roth, B. Schiele, People-tracking-by-detection and people-detection-by-tracking, in: Proceedings of the Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[16] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of the International Conference on Machine Learning, 2007, pp. 209–216.
[17] M. Dikmen, E. Akbas, T.S. Huang, N. Ahuja, Pedestrian recognition with a learned metric, Lect. Notes Comput. Sci. 6495 (2011) 501–512.
[18] J. Lai, J. Lai, J. Lai, Mirror representation for modeling view-specific transform in person re-identification, in: Proceedings of the International Conference on Artificial Intelligence, 2015, pp. 3402–3408.
[19] X. Wang, W.S. Zheng, X. Li, J. Zhang, Cross-scenario transfer person reidentification, IEEE Trans. Circuits Syst. Video Technol. 26 (2016) 1447–1460.
[20] S. Khamis, C. Kuo, V.K. Singh, V.D. Shet, L.S. Davis, Joint learning for attribute–consistent person re-identification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 134–146.
[21] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person Re-Identification Using Kernel-Based Metric Learning Methods, Springer International Publishing, 2014.
[22] R. Satta, Appearance descriptors for person re-identification: a comprehensive review, in: Proceedings of the Computer Vision and Pattern Recognition, 2013.

[23] L. Zheng, Y. Yang, A.G. Hauptmann, Person re-identification: past, present and future, in: Proceedings of the Computer Vision and Pattern Recognition, 2016.
[24] T. D'Orazio, G. Cicirelli, People re-identification and tracking from multiple cameras: a review, in: Proceedings of the IEEE International Conference on Image Processing, 2013, pp. 1601–1604.
[25] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification ☆, Image Vis. Comput. 32 (2014) 270–286.
[26] R. Satta, Appearance Descriptors for Person Re-identification: A Comprehensive Review[J], arXiv: Computer Vision and Pattern Recognition (2013).
[27] X. Chu, W. Ouyang, H. Li, X. Wang, Structured feature learning for pose estimation, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 4715–4723.
[28] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[29] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the Neural Information Processing Systems, 2012, pp. 1097–1105.
[30] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[31] W. Yang, W. Ouyang, H. Li, X. Wang, End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 3073–3082.
[32] Z.Q. Zhao, D.S. Huang, B.Y. Sun, Human face recognition based on multi-features using neural networks committee, Pattern Recognit. Lett. 25 (2004) 1351–1358.
[33] D.S. Huang, A constructive approach for finding arbitrary roots of polynomials by neural networks, IEEE Trans. Neural Netw. 15 (2004) 477.
[34] L. Shang, D.S. Huang, J.X. Du, C.H. Zheng, Palmprint recognition using FastICA algorithm and radial basis probabilistic neural network ☆, Neurocomputing 69 (2006) 1782–1786.
[35] D.S. Huang, H.H.S. Ip, Z. Chi, A neural root finder of polynomials based on root moments, Neural Comput. 16 (2004) 1721–1762.
[36] D.-S. Huang, Radial basis probabilistic neural networks: model and application, Int. J. Pattern Recognit. Artif. Intell. 13 (1999) 1083–1101.
[37] X.F. Wang, D.S. Huang, A novel density-based clustering framework by using level set method, IEEE Trans. Knowl. Data Eng. 21 (2009) 1515–1531.
[38] K. Kang, W. Ouyang, H. Li, X. Wang, Object detection from video tubelets with convolutional neural networks, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 817–825.
[39] B. Li, C.-H. Zheng, D.-S. Huang, Locally linear discriminant embedding: an efficient method for face recognition, Pattern Recognit. 41 (2008) 3813–3821.
[40] D.-S. Huang, J.-X. Du, A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks, IEEE Trans. Neural Netw. 19 (2008) 2099–2115.
[41] X.F. Wang, D.S. Huang, H. Xu, An efficient local Chan–Vese model for image segmentation, Pattern Recognit. 43 (2010) 603–618.
[42] D.S. Huang, W. Jiang, A general CPL-AdS methodology for fixing dynamic parameters in dual environments, IEEE Trans. Syst. Man Cybern. Part B Cybern. A Publ. IEEE Syst. Man Cybern. Soc. 42 (2012) 1489–1500.
[43] W. Zhao, D. Huang, J. Du, L. Wang, Genetic optimization of radial basis probabilistic neural networks, Int. J. Pattern Recognit. Artif. Intell. 18 (2008) 1473–1499.
[44] D.S. Huang, S.D. Ma, Linear and nonlinear feedforward neural network classifiers: a comprehensive understanding, J. Intell. Syst. 9 (1999) 1–38.
[45] Z.Q. Zhao, D.S. Huang, A mended hybrid learning algorithm for radial basis function neural networks to improve generalization capability ☆, Appl. Math. Model. 31 (2007) 1271–1281.
[46] J.X. Du, D.S. Huang, G.J. Zhang, Z.F. Wang, A novel full structure optimization algorithm for radial basis probabilistic neural networks, Neurocomputing 70 (2006) 592–596.
[47] F. Han, D.S. Huang, A new constrained learning algorithm for function approximation by encoding a priori information into feedforward neural networks, Neural Comput. Appl. 17 (2008) 433–439.
[48] J.X. Du, D.S. Huang, X.F. Wang, X. Gu, Shape recognition based on neural networks trained by differential evolution algorithm, Neurocomputing 70 (2007) 896–903.
[49] F. Han, Q.H. Ling, D.S. Huang, Modified Constrained Learning Algorithms Incorporating Additional Functional Constraints Into Neural Networks, Elsevier Science Inc, 2008.
[50] D.S. Huang, W.B. Zhao, Determining the centers of radial basis probabilistic neural networks by recursive orthogonal least square algorithms, Appl. Math. Comput. 162 (2005) 461–473.
[51] D.S. Huang, H.H.S. Ip, K.C.K. Law, Z. Chi, Zeroing polynomials using modified constrained neural network approach, IEEE Trans. Neural Netw. 16 (2005) 721–732.
[52] Z.Q. Zhao, D.S. Huang, W. Jia, Palmprint recognition with 2DPCA+PCA based on modular neural networks, Neurocomputing 71 (2007) 448–454.
[53] D.S. Huang, J.X. Mi, A new constrained independent component analysis method, IEEE Trans. Neural Netw. 18 (2007) 1532–1535.
[54] S. Wu, Y.C. Chen, X. Li, A.C. Wu, J.J. You, W.S. Zheng, An enhanced deep feature representation for person re-identification, in: Proceedings of the Applications of Computer Vision, 2016, pp. 1–8.
[55] L. Wu, C. Shen, A.V.D. Hengel, Deep linear discriminant analysis on fisher networks: a hybrid architecture for person re-identification, Pattern Recognit. 65 (2016) 238–250.
[56] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.
[57] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A Discriminative Feature Learning Approach For Deep Face Recognition, Springer International Publishing, 2016.
[58] H. Jin, X. Wang, S. Liao, et al., Deep person re-identification with improved embedding and efficient training[J], International Journal of Central Banking (2017) 261–267.
[59] L. Zheng, Y. Huang, H. Lu, et al., Pose Invariant Embedding for Deep Person Re-identification. [J], arXiv: Computer Vision and Pattern Recognition (2017).
[60] Z. Zheng, L. Zheng, Y. Yang, et al., Pedestrian Alignment Network for Large-scale Person Re-identification[J], IEEE Transactions on Circuits and Systems for Video Technology (2018) 1-1.
[61] Y. Lin, L. Zheng, Z. Zheng, et al., Improving Person Re-identification by Attribute and Identity Learning. [J], arXiv: Computer Vision and Pattern Recognition (2017).
[62] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, et al., Improving deep visual representation for person re-identification by global and local image-language association, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 56–73.
[63] L. Zheng, H. Zhang, S. Sun, et al., Person Re-identification in the Wild[J], Computer vision and pattern recognition (2017) 3346–3355.
[64] T. Xiao, S. Li, B. Wang, et al., End-to-End Deep Learning for Person Search. [J], arXiv: Computer Vision and Pattern Recognition (2016).
[65] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3376–3385.
[66] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[67] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: Proceedings of the Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.
[68] L. Wu, C. Shen, A.V. Den Hengel, et al., PersonNet: Person Re-identification with Deep Convolutional Neural Networks[J], arXiv: Computer Vision and Pattern Recognition (2016).
[69] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: Proceedings of the Twenty-Second International Conference on Pattern Recognition (ICPR), 2014, pp. 34–39.
[70] Y. Chen, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of the International Conference on Neural Information Processing Systems, 2014, pp. 1988–1996.
[71] Z. Zheng, L. Zheng, Y. Yang, et al., A Discriminatively Learned CNN Embedding for Person Reidentification[J], ACM Transactions on Multimedia Computing, Communications, and Applications 14 (1) (2017).
[72] H. Chen, Y. Wang, Y. Shi, et al., Deep Transfer Learning for Person Re-Identification[J], IEEE international conference on multimedia big data (2018) 1–5.
[73] X. Qian, Y. Fu, Y. Jiang, et al., Multi-scale Deep Learning Architectures for Person Re-identification[J], International conference on computer vision (2017) 5409–5418.
[74] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, et al., Learning fine-grained image similarity with deep ranking, in: Proceedings of the Computer Vision and Pattern Recognition, 2014, pp. 1386–1393.
[75] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[76] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognit. 48 (2015) 2993–3003.
[77] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.
[78] A. Hermans, L. Beyer, B. Leibe, et al., In Defense of the Triplet Loss for Person Re-Identification. [J], arXiv: Computer Vision and Pattern Recognition (2017).
[79] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 475–491.
[80] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification, IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc. PP (2016) 1.
[81] X. Bai, M. Yang, T. Huang, et al., Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. [J], arXiv: Computer Vision and Pattern Recognition (2017).
[82] G. Wang, Y. Yuan, X. Chen, et al., Learning Discriminative Features with Multiple Granularities for Person Re-Identification. [J], Acm multimedia (2018) 274–282.
[83] W. Chen, X. Chen, J. Zhang, K. Huang, A multi-task deep network for person *re*-identification, in: Proceedings of the Association for the Advancement of Artificial Intelligence AAAI, 2017, pp. 3988–3994.
[84] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: Proceedings of the Computer Vision and Pattern Recognition, 2016, pp. 1288–1296.

[85] R.R. Varior, M. Haloi, G. Wang, Gated Siamese convolutional neural network architecture for human re-identification, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 791–808.
[86] W. Chen, X. Chen, J. Zhang, et al., Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification[J], Computer vision and pattern recognition (2017) 1320–1329.
[87] Q. Xiao, H. Luo, C. Zhang, et al., Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification. [J], arXiv: Computer Vision and Pattern Recognition (2017).
[88] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, X. Bai, Hard-aware point-to-set deep metric for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 196–212.
[89] H. Shi, Y. Yang, X. Zhu, et al., Embedding Deep Metric for Person Re-identification: A Study Against Large Variations[J], European conference on computer vision (2016) 732–748.
[90] E. Ustinova, Y. Ganin, V. Lempitsky, Multi-region bilinear convolutional neural networks for person re-identification, in: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017, pp. 2993–3003.
[91] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A Siamese long short-term memory architecture for human re-identification, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 135–153.
[92] W. Li, X. Zhu, S. Gong, et al., Person Re-Identification by Deep Joint Learning of Multi-Loss Classification[J], International joint conference on artificial intelligence (2017) 2194–2200.
[93] D. Li, X. Chen, Z. Zhang, et al., Learning Deep Context-Aware Features over Body and Latent Parts for Person Re-identification[J], Computer vision and pattern recognition (2017) 7398–7407.
[94] C. Su, J. Li, S. Zhang, et al., Pose-Driven Deep Convolutional Model for Person Re-identification[J], International conference on computer vision (2017) 3980–3989.
[95] H. Yao, S. Zhang, R. Hong, et al., Deep representation learning with part loss for person re-identification[J], IEEE Transactions on Image Processing (2019).
[96] Y. Suh, J. Wang, S. Tang, T. Mei, K.M. Lee, Part-aligned bilinear representations for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 402–419.
[97] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision, 2018, pp. 480–496.
[98] W. Li, X. Zhu, S. Gong, et al., Harmonious Attention Network for Person Re-Identification[J], Computer vision and pattern recognition (2018) 2285–2294.
[99] X. Liu, H. Zhao, M. Tian, et al., HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis[J], International conference on computer vision (2017) 350–359.
[100] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: a multi-task attentional network with curriculum sampling for person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 384–400.
[101] L. Zhao, X. Li, Y. Zhuang, et al., Deeply-Learned Part-Aligned Representations for Person Re-identification[J], International conference on computer vision (2017) 3239–3248.
[102] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, et al., Mars: a video benchmark for large-scale person re-identification, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 868–884.
[103] N. Mclaughlin, J.M.D. Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1325–1334.
[104] L. Wu, C. Shen, A.V. Den Hengel, et al., Deep Recurrent Convolutional Networks for Video-based Person Re-identification: An End-to-End Approach. [J], arXiv: Computer Vision and Pattern Recognition (2016).
[105] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, X. Yang, Person re-identification via recurrent feature aggregation, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 701–716.
[106] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, et al., Video-based person re-identification with accumulative motion context, IEEE Trans. Circuits Syst. Video Technol. PP (2017) 1.
[107] Y. Li, L. Zhuo, J. Li, J. Zhang, X. Liang, Q. Tian, Video-based person re-identification by deep feature guided pooling, in: Computer Vision and Pattern Recognition Workshops, 2017, pp. 1454–1461.
[108] D. Chung, K. Tahboub, E.J. Delp, A two stream Siamese convolutional neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1983–1991.
[109] S. Xu, Y. Cheng, K. Gu, et al., Jointly Attentive Spatial-Temporal Pooling Networks for Video-Based Person Re-identification[J], International conference on computer vision (2017) 4743–4752.
[110] Y. Shen, H. Li, S. Yi, et al., Person Re-identification with Deep Similarity-Guided Graph Neural Network[C], in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 486–504.
[111] Y. Chen, X. Zhu, S. Gong, Deep association learning for unsupervised video person re-identification, in: Proceedings of the British Machine Vision Conference, 2018, p. 48.
[112] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro[J], arXiv preprint arXiv 1701 (07717) (2017) 3.
[113] Z. Zhong, L. Zheng, Z. Zheng, et al., Camera style adaptation for person re-identification[C], in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5157–5166.
[114] L. Wei, S. Zhang, W. Gao, et al., Person transfer gan to bridge domain gap for person re-identification[C], in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 79–88.
[115] X. Qian, Y. Fu, T. Xiang, et al., Pose-normalized image generation for person re-identification[C], in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 650–667.
[116] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2242–2251.
[117] R. Panda, A. Bhuiyan, V. Murino, et al., Unsupervised Adaptive Re-identification in Open World Dynamic Camera Networks[J], Computer vision and pattern recognition (2017) 1377–1386.
[118] H. Fan, L. Zheng, C. Yan, et al., Unsupervised person re-identification: Clustering and fine-tuning[J], ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 (4) (2018) 83.
[119] J. Wang, X. Zhu, S. Gong, et al., Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification[J], Computer vision and pattern recognition (2018) 2275–2284.
[120] W. Deng, L. Zheng, Q. Ye, et al., Image-Image Domain Adaptation With Preserved Self-Similarity and Domain-Dissimilarity for Person Re-Identification[J], Computer vision and pattern recognition (2018) 994–1003.
[121] S. Lin, H. Li, C.T. Li, A.C. Kot, Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification, in: Proceedings of the British Machine Vision Conference, 2018, p. 9.
[122] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero- and homogeneously, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 172–188.
[123] J. Lin, L. Ren, J. Lu, J. Feng, J. Zhou, Consistent-aware deep learning for person re-identification in a camera network, in: Proceedings of the Computer Vision and Pattern Recognition, 2017, pp. 3396–3405.
[124] X. Li, A. Wu, W. Zheng, Adversarial open-world person re-identification, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 287–303.
[125] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, F. Porikli, Feature affinity based pseudo labeling for semi-supervised person re-identification, in: Proceedings of the Computer Vision and Pattern Recognition, 2018.
[126] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: Proceedings of the International Conference on Computer Vision, 2017, pp. 3774–3782.
[127] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, J. Zhang, Multi-pseudo regularized label for generated data in person re-identification, IEEE Trans. Image Process. 28 (2018) 1391–1403.
[128] Z. Wang, M. Ye, F. Yang, X. Bai, S. Satoh, Cascaded SR-GAN for scale-adaptive low resolution person re-identification, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2018, pp. 3891–3897.
[129] J. Jiao, W. Zheng, A. Wu, X. Zhu, S. Gong, Deep low-resolution person re-identification, in: Proceedings of the National Conference on Artificial Intelligence, 2018, pp. 6967–6974.
[130] F. Zhu, X. Kong, L. Zheng, H. Fu, Q. Tian, Part-based deep hashing for large-scale person re-identification, IEEE Trans. Image Process. 26 (2017) 4806–4817.
[131] L. Wu, Y. Wang, Z. Ge, Q. Hu, X. Li, Structured deep hashing with convolutional neural networks for fast person re-identification, Comput. Vis. Image Underst. 167 (2017) 63–73.
[132] A. Haque, A. Alahi, L. Feifei, Recurrent attention models for depth-based person identification, Comput. Vis. Pattern Recognit. (2016) 1229–1238.
[133] F. Pala, R. Satta, G. Fumera, F. Roli, Multimodal person reidentification using RGB-D cameras, IEEE Trans. Circuits Syst. Video Technol. 26 (2016) 788–799.
[134] P. Dai, R. Ji, H. Wang, Q. Wu, Y. Huang, Cross-modality person re-identification with generative adversarial training, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2018, pp. 677–683.
[135] M. Ye, Z. Wang, X. Lan, P.C. Yuen, Visible thermal person re-identification via dual-constrained top-ranking, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2018, pp. 1092–1099.
[136] A. Wu, W. Zheng, H. Yu, S. Gong, J. Lai, RGb-infrared cross-modality person re-identification, in: Proceedings of the International Conference on Computer Vision, 2017, pp. 5390–5399.
[137] F. Hafner, A. Bhuiyan, J.F.P. Kooij, et al., A Cross-Modal Distillation Network for Person Re-identification in RGB-Depth[J], arXiv preprint arXiv 1810 (2018) 11641.
[138] M. Munaro, A. Basso, A. Fossati, L. Van Gool, E. Menegatti, 3D reconstruction of freely moving persons for re-identification with a depth sensor, in: Proceedings of the International Conference on Robotics and Automation, 2014, pp. 4512–4519.
[139] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), 3, 2007, pp. 1–7.
[140] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Proceedings of the Scandinavian Conference on Image Analysis, 2011, pp. 91–102.

[141] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2016, pp. 1116–1124.
[142] C.L. Chen, T. Xiang, S. Gong, Multi-camera activity correlation analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, 2009, pp. 1988–1995.
[143] S. Karanam, M. Gou, Z. Wu, et al., A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets, IEEE Trans. Pattern Anal. Mach. Intell. (2018) 1.
[144] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, et al., From the lab to the real world: re-identification in an airport camera network, IEEE Trans. Circuits Syst. Video Technol. 27 (2017) 540–553.
[145] W.S. Zheng, L. Xiang, X. Tao, S. Liao, J. Lai, S. Gong, Partial person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2016.
[146] A. Ess, B. Leibe, L.V. Gool, Depth and appearance for mobile scene analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[147] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 688–703.
[148] D. Baltieri, R. Vezzani, R. Cucchiara, 3DPeS: 3D people dataset for surveillance and forensics, in: Proceedings of the Joint ACM Workshop on Human Gesture and Behavior Understanding, 2011, pp. 59–64.
[149] L. Zheng, Z. Bie, Y. Sun, et al., Mars: A video benchmark for large-scale person re-identification[C], European Conference on Computer Vision, Springer, Cham, 2016, pp. 868–884.
[150] I.B. Barbosa, M. Cristani, A.D. Bue, L. Bazzani, V. Murino, Re-identification with RGB-D sensors, in: Proceedings of the International Conference on Computer Vision, 2012, pp. 433–442.
[151] M. Munaro, A. Fossati, A. Basso, et al., One-shot person re-identification with a consumer depth camera[M], Person Re-Identification, Springer, London, 2014, pp. 161–181.
[152] M. Hirzer, Large scale metric learning from equivalence constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[153] W. Zheng, S. Gong, T. Xiang, et al., Associating groups of people, british machine vision conference, 2009, pp. 1–11.
[154] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the International Conference on Computer Vision, 2015, pp. 3730–3738.

Di Wu is now a Ph.D. candidate with the Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, China. He received the B.S. degree from Zhengzhou University, China, in 2013, and the M.S. degree from Zhengzhou Institute of Light Industry, China, in 2016. His research focuses on deep learning and image processing.

Sijia Zheng is now a Master candidate with the Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, China. She received the B.S. degree from Shandong University, China, in 2016. Her research focuses on deep learning and image processing.

Xiao-Ping (Steven) Zhang received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in electronic engineering and the M.B.A. degree in finance, economics, and entrepreneurship from the University of Chicago, Illinois. He is a professor of electrical and computer engineering and is cross appointed to the Finance Department at the Ted Rogers School of Management at Ryerson University, Toronto, Canada. His research interests include signal processing, electronic systems, machine learning, big data, finance, and marketing. He is the cofounder and chief executive officer for EidoSearch, an Ontario-based company offering a content-based search and analysis engine for financial big data.

De-Shuang Huang received the B.Sc., M.Sc. and Ph.D. degrees all in electronic engineering from Institute of Electronic Engineering, Hefei, China, National Defense University of Science and Technology, Changsha, China and Xidian University, Xian, China, in 1986, 1989 and 1993, respectively. During 1993–1997 periods, he was a postdoctoral research fellow respectively in Beijing Institute of Technology and in National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In Sept, 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences as the Recipient of "Hundred Talents Program of CAS". In September 2011, he entered into Tongji University as Chaired Professor. From Sept 2000 to Mar 2001, he worked as Research Associate in Hong Kong Polytechnic University. From Aug. to Sept. 2003, he visited the George Washington University as visiting professor, WashingtonDC, USA. From July to Dec 2004, he worked as the University Fellow in Hong Kong Baptist University. From March, 2005 to March, 2006, he worked as Research Fellow in Chinese University of Hong Kong. From March to July, 2006, he worked as visiting professor in Queen's University of Belfast, UK. In 2007, 2008, 2009, he worked as visiting professor in Inha University, Korea, respectively. At present, he is the director of Institute of Machines Learning and Systems Biology, Tongji University. He is currently Fellow of International Association of Pattern Recognition (IAPR Fellow), senior members of the IEEE and International Neural Networks Society. He has published over 180 journal papers. Also, in 1996, he published a book entitled "Systematic Theory of Neural Networks for Pattern Recognition" (in Chinese), which won the Second-Class Prize of the 8th Excellent High Technology Books of China, and in 2001 and 2009 another two books entitled "Intelligent Signal Processing Technique for High Resolution Radars" (in Chinese) and "The Study of Data Mining Methods for Gene Expression Profiles" (in Chinese), respectively. His current research interest includes bioinformatics, pattern recognition and machine learning.