CAAI Transactions on Intelligence Technology

Review Article

Survey on person re-identification based on deep learning

ISSN 2468-2322
Received on 12th June 2018
Accepted on 12th June 2018
doi: 10.1049/trit.2018.1001
www.ietdl.org

Kejun Wang[1], Haolin Wang[1], Meichen Liu[1], Xianglei Xing[1] ✉, Tian Han[2]

[1]*College of Automation, Harbin Engineering University, No. 145 Nantong Street, Nangang District, Harbin, People's Republic of China*
[2]*Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095, USA*
✉ *E-mail: xingxl@hrbeu.edu.cn*

Abstract: Person re-identification (Re-ID) is a fundamental subject in the field of the computer vision technologies. The traditional methods of person Re-ID have difficulty in solving the problems of person illumination, occlusion and attitude change under complex background. Meanwhile, the introduction of deep learning opens a new way of person Re-ID research and becomes a hot spot in this field. This study reviews the traditional methods of person Re-ID, then the authors focus on the related papers about different person Re-ID frameworks on the basis of deep learning, and discusses their advantages and disadvantages. Finally, they propose the direction of further research, especially the prospect of person Re-ID methods based on deep learning.

1 Introduction

The camera network is increasingly deployed in public places like airports, railway stations, college campuses and office buildings. These cameras typically span large geospatial areas and have non-overlapping fields-of-views to provide enhanced coverage. These cameras provide a huge amount of video data which is manually monitored by law enforcement officers. Therefore, people need to analyse these video data through the computer vision technologies. Person re-identification (Re-ID) has become increasingly popular in the community due to its application and research significance. Person Re-ID aims to match the person across multiple camera views. Fig. 1 illustrates an example of a surveillance area monitored by multiple cameras with non-overlapping fields-of-views. It is attracting rapidly increased attentions in the computer vision and pattern recognition research community due to its importance for many applications such as video surveillance, human–computer interaction, robotics, content-based video retrieval and so on.

Person Re-ID still has many challenges, such as large variations in person pose, illumination, and background clutter. In addition, similar appearance of clothes among different people and imperfect pedestrian detection results further increase its difficulty in real applications. Deep learning was proposed by Hinton and Salakhutdinov [2], and its adaptability is good at mining deep features of data and has achieved good results in the fields of image classification, speech recognition and visual tracking. Recently, deep learning approaches have achieved the state-of-the-art results for person Re-ID. Liu *et al.* [3] proposed a novel Accumulative Motion Context (AMOC) network. Compared to traditional methods, such as Top-push Distance Learning Model (TDL) [4] and spatio-temporal appearance (STA) [5], the AMOC network can improve Rank-1 identification rate by 13% in iLIDS-VID [6] datasets. Li *et al.* [7] designed a Multi-Scale Context-Aware Network (MSCAN) to learn the powerful features over the full body and body parts. The MSCAN improved Rank-1 identification rate by >35% compared with Bag-of-Words Model (BOW) [8] and Weighted Approximate Rank Component Analysis (WARCA) [9] on the Market-1501 [8]. Bak and Carr [10] proposed that the combination of deep learning and traditional method achieved Rank-1 identification rate by 41.4% on the PRID [11] dataset, but traditional method local maximal occurrence (LOMO)+Cross-view Quadratic Discriminant Analysis (XQDA) [12] only achieved Rank-1 identification rate by 26.7% on this dataset.

Person Re-ID based on deep learning models has been a hot research spot. In order to attract more researchers in the field of person Re-ID to explore and discuss deep learning, and to promote the research of person Re-ID algorithms, this paper reviews the traditional designs of person Re-ID methods. Then we mainly introduce the research methods of person re-recognition based on deep learning till the end of 2017 and summarise advantages and disadvantages of various algorithms. Most of the algorithms are published in international conferences such as IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV) and other famous periodicals. These research ideas and performances are representative and innovative. Finally, we propose the further research directions and the prospect of person Re-ID method based on deep learning.

2 Typical person Re-ID methods

The early person Re-ID can be traced back to 1997. After 2008, person Re-ID has been developing rapidly. Typical person Re-ID methods can be roughly divided into two categories: feature-based and metric-based. Feature-based methods focus on seeking effective descriptors for person representation. Metric-based methods focus on learning an effective metric to reduce the distance of same people and increase the distance of different people. The following will introduce the traditional person Re-ID methods from these aspects, including the classic methods and the new and innovative algorithms.

2.1 Feature-based methods

Due to the poor resolution of the images in the surveillance video and unconstrained acquisition environment, we cannot rely on face recognition technology to achieve person Re-ID. On the basis of person representation information, many research designed stable and discriminative feature descriptors, including colour features, texture features, local features and semantic features. Algorithm diagram is shown in Fig. 2.

Zhao *et al.* [13] used the colour histogram and scale-invariant feature transform (SIFT) [14] to extract image feature and applied adjacency constrained patch matching to build a correspondence between image pairs. This method shows effectiveness in handling

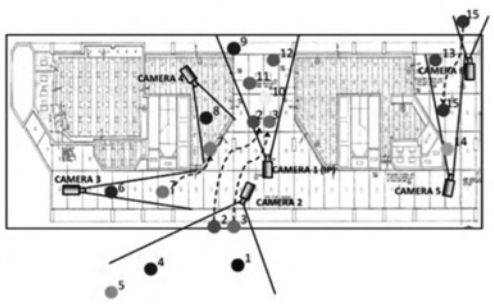

Fig. 1 *Multi-camera surveillance network illustration of Re-ID [1]*

Fig. 2 *Re-ID system based on feature-based methods*

misalignment caused by the viewpoint and pose variations. Shen and co-authors [15] used SIFT [14] and LAB colour histogram to extract the features as a characteristic representation of a person.

Most of the feature descriptors extract the global image features, such as colour, texture and shape. These features are not robust in uncalibrated images and lack space constraints. To solve this problem, the authors of [16–18] adopted the local feature matching methods. Chen *et al.* [16] put forward to a block matching method via the colour distribution fields with adaptive hierarchical structure. Diving the person image into multiply overlapping pieces can solve the accuracy problem of the body parts and make use of the finer granularity information. Yang *et al.* [17] divided the body into three parts by using Symmetry-Driven Accumulation of Local Features (SDALF) and then extracted the Hue, Saturation, Value (HSV) histogram and texture features of each block. Geng *et al.* [18] proposed to fully exploit region-based feature salience. Firstly, a part-based feature extraction algorithm was proposed to adopt different features for different parts correspondingly. Secondly, the salient colour descriptor was proposed by considering the colour diversity between the current region and its surrounding regions.

In order to effectively reduce the variation of the cross-viewing angles, Layne *et al.* [19] extracted semantic features for person Re-ID, marking the characteristics of a person with 15 semantic attributes (such as dress, skin colour etc.). Then they merged low-level visual characteristics with semantic attributes on the basis of weighting different attributes. However, it is difficult to understand the attribute characters of person images in grainy video footage, so there are few research results in the semantic properties aspect.

Most existing typical feature extraction methods are focusing on describing person characteristic. However, they still have many challenges when considering reality applications, such as large variations in person pose, background clutter and foreground cover. They make the inter-class difference much more obscure and feature less discriminative.

2.2 Metric-based methods

The good distance metric is critical for Re-ID systems' success because the high-dimensional visual features typically do not capture the invariant factors from various samples. Once the feature extraction is completed, these methods usually choose a standard distance measure such as Mahalanobis distance to determine the similarity between pairs of images. The loss function is used to train the models, which makes the distance between the matched pairs less than the mismatched pairs in the learned feature space. The general idea of metric learning is to keep all the vectors of the same class closer while pushing vectors of different classes further apart. Algorithm diagram is shown in Fig. 3. The most common formulation is on the basis of Mahalanobis distance function, which generalises Euclidean distance by using linear scalings and rotations in the feature space.

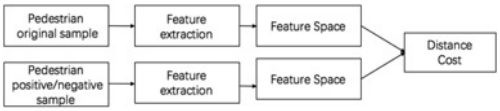

Fig. 3 *Re-ID system based on metric-based methods*

When using video-based representation, some inter-class differences can be much more obscure than the ones when using still-image based representation, because different people could not only have a similar appearance but also have similar motions and actions which are hard to align. To solve this problem, You *et al.* [4] proposed a Top-push Distance Learning (TDL) based on LMNN [20], in which they integrated a top-push constrain for matching video features of persons. The top-push constraint enforces the optimisation on top-rank matching in Re-ID, so as to make the matching model more effective by selecting more discriminative features to distinguish different persons. Yu *et al.* [21] proposed unsupervised asymmetric metric learning for unsupervised Re-ID. The model aims to learn an asymmetric metric, i.e. the specific projection for each view, based on asymmetric clustering on cross-view person images. Zhong *et al.* [22] proposed a k-reciprocal encoding method to re-rank the Re-ID results. Given an image, a k-reciprocal feature is calculated by encoding its k-reciprocal nearest neighbours into a single vector, which is used for re-ranking under the Jaccard distance. The final distance is computed as the combination of the original distance and the Jaccard distance. This method does not require any human interaction or any labelled data, so it is applicable for large-scale datasets. Chen *et al.* [23] proposed Cross-camera Semantic Binary Transformation (CSBT) method. CSBT employs a subspace projection to mitigate cross-camera variations by maximising intra-person similarities and inter-person discrepancies. And a binary coding scheme was proposed via seamlessly incorporating both the pair-wise semantic relationships and local affinity information. It is proposed for sub-space projection learning and binary coding based on discrete alternating optimisation. CSBT aims to transform original high-dimensional feature vectors into compact identity-preserving binary codes.

The metric-based methods are largely limited by feature representation and robustness. The distance measure obtained by supervised learning can only be applied to a particular scenario, and it is less effective when migrating to a new environment.

Most existing traditional methods focus on designing features manually and cannot capture advanced semantic information. Compared with the traditional methods, deep learning has strong abilities of independent learning and feature extraction. In order to enhance the generalisation ability of the models, many scholars use deep learning to achieve person Re-ID by avoiding the loss of information in the process of feature extraction.

3 Person Re-ID based on deep learning

The success of deep learning in image classification spread to person Re-ID in 2014, when Li *et al.* [24] and Yi *et al.* [25] both employed a Siamese neural network to determine if a pair of input images belonged to the same ID. The Rank-1 recognition rate reached 27.87% by Li *et al.* [24] on the CUHK01 dataset. The Yi *et al.* [25] method reached Rank-1 recognition rate by 23.23% on the VIPeR dataset. Both of them had outperformed the traditional methods at that time. At present, there are many methods for person Re-ID on the basis of deep learning. A large number of academic papers have been published in mainstream conferences and international authoritative publications such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), International Journal of Conflict and Violence (IJCV), IEEE Transactions on Image Processing (TIP). The following will introduce the mainstream algorithms of person Re-ID based on deep learning, and especially introduce the latest and innovate research in details. According to the deep learning model, person Re-ID methods are divided into three parts: convolution neural

Fig. 4 *Re-ID system based on CNN diagram*

network (CNN)-based methods, person Re-ID based on CNN and recursive neural network (RNN), generative adversarial network (GAN)-based methods and the hybrid of deep learning and traditional method.

3.1 CNN-based methods

CNN origins from a multi-layer perceptron with two-dimensional shape inspired by the visual system of neural mechanisms. Lecun *et al.* [26] took a combination of convolutional layer and pooling layer to handwritten character recognition task, and the obvious effect was obtained. The CNN is used to achieve the advantages of rotation invariance, position invariance and scale invariance, which can extract the person detailed features more accurately. The CNN-based person Re-ID methods use only CNN model, which combined with different networks and matching algorithms. Algorithm diagram is shown in Fig. 4. CNN-based person Re-ID methods can be divided into three categories according to the types of image information extracted by CNN. It includes local-based network structure, multi-scale network structure, and multi-model network structure.

3.1.1 Local-based network structure: The person Re-ID appearance is often influenced by some factors, such as pose, illumination and so on, which means same person generally have different appearances. Li *et al.* [27] added a horizontal region matching layer and used two filters to match the regions. Cheng *et al.* [28] designed a triplet loss function that takes three images as inputs. After the first convolution layer, four overlapping body parts were partitioned for each image and fused with a global one in the full connect layer. Considering that the person videos mainly come from the monitoring data with a lower resolution, the local area size becomes smaller after segmentation. In order to make extracted features as comprehensive as possible, Li and Chen [29] introduced the deformable part model [30] person segmentation method. They sent each segmented region into the CNN model with the spatial pyramid pooling layer, and then achieved the person Re-ID on the basis of fusing the recognition results of each part. Varior *et al.* [31] proposed to add a gate function after each convolution layer to calculate the attention scores of each local area in the image. By introducing the concept of the attention score, the method can capture the subtle and effective differences.

3.1.2 Multi-scale network structure: CNN can effectively extract the target features from the spatial information. However, considering person image variability and its complex background, it is impossible to extract discriminative the features by using convolutional layer networks alone. To solve this problem, scholars have proposed a variety of network structure algorithms based on multi-scale image information. Ahmed *et al.* [32] imported a pair of person images into the CNN to extract features. Finally, a softmax function yields the final estimate whether the input images come from the same person or not. Jiang and co-authors [33] proposed a novel multi-scale deep learning model. The model is able to learn deep discriminative feature representations at different scales and automatically determined the most suitable scales for matching. Liu *et al.* [34] proposed a new attention-based CNN, named as HydraPlus-Net (HP-net) that multi-directionally feeds the multi-level attention maps to different feature layers. As shown in Fig. 5, the HP-net comprises the Main Net (M-net) and the Attentive Feature Net (AF-net). M-net is a plain CNN architecture. AF-net uses multiple branches of multi-directional attention modules to extract semantic features on different levels. Their outputs are concatenated and then fused by global average pooling and fully connected layers. The model is

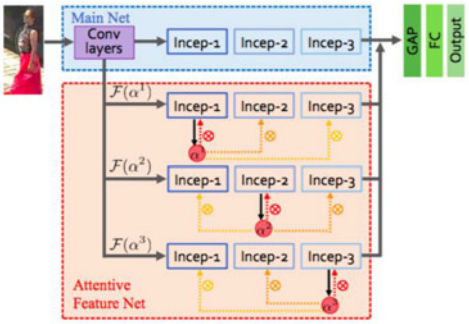

Fig. 5 *HP-Net with a main net and an attentive feature net [34]*

capable of capturing multiple attentions from low level to semantic level, and it explores the multi-scale selectiveness of attentive features to enrich the final feature representations of a person image. However, Li *et al.* [7] designed a Multi-Scale Context-Aware Network (MSCAN) to learn the powerful features over full body and body parts, which can capture the local context knowledge by stacking multi-scale convolutions in each layer. As shown in Fig. 6, the model comprises three components: the global body-based feature learning with MSCAN, the latent pedestrian parts localisation with Spatial Transformer Networks [36] and local part-based feature embedding, the fusion of full body and body parts for multi-class person Re-ID tasks. The algorithm achieved Rank-1 identification rate by 80.31% on the Market-1501 dataset. In addition, human body dislocation caused by detector or pose change is sometimes serious for across images feature matching. Zheng *et al.* [37] proposed a convolution network method called PoseBox [38], which constructs a PoseBox for each person. PoseBox generates standard alignment images by pose estimation and affine projection. They sent the PoseBox to the ResNet-50 [39] network with the original image as inputs for training. The advantage of this algorithm is that it can overcome the deformation problem caused by person pose changing and still finds the same person when the posture changes. Su *et al.* [40] also proposed a deep convolution model called PDC, which extracts body part features from person pose information, and trains two parallel convolution neural networks with a new triple loss function. This method reduces the influence of posture diversity on Re-ID effect. Zhao *et al.* [41] proposed a Spindle Net convolutional network on the basis of human body structure information, which consists of Feature Extraction Network (FEN) and Feature Fusion Network (FFN). The FEN network mainly takes person images and seven body regions extracted by the Region Proposal Network [42] as inputs, and calculate the global feature vector of the whole image, as well as the vectors corresponding to the seven sub-regions. Through the FFN network, these feature vectors are combined to the final feature vector, which is used for distinguishing people. The flowchart is shown in Fig. 7. The advantages of this algorithm are that it can

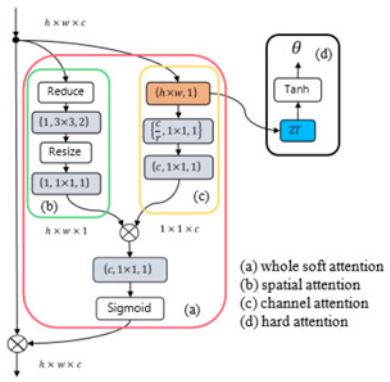

Fig. 6 *Structure of each harmonious attention [35] module*

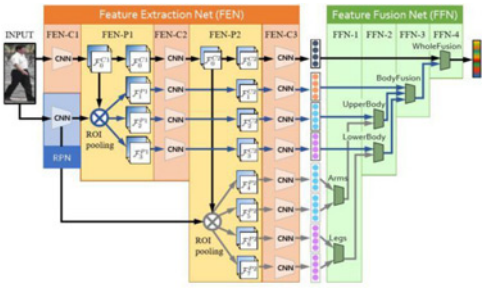

Fig. 7 *Spindle Net [41]. This architecture includes the FEN and the FFN*

preserve and capture semantic features from different body regions. Li *et al.* [35] proposed Hierarchical-Associative convolution neural network (HA-CNN) model, which uses the Inception module to construct four parallel networks, extracting image features from multiple scales. Four networks share weights in some modules, which reduce the parameter size. Based on this, the HA module (Fig. 6) is used to extract discriminative features with attention modules. It includes four kinds of attention structures: soft attention, spatial attention, channel attention, and hard regional attention.

3.1.3 Multi-model network structure: Datasets are critical for testing the feasibility of person Re-ID system. And they are not large enough to provide abundant data variations. In order to let the neural network learn more patterns from images and improve its generalisation performance, Xiao *et al.* [43] presented a pipeline for learning deep feature representations from multiple domains with CNN. In [43], training identities from multiple datasets unite and a new softmax loss based on softmax is employed in the classification network. Together with the proposed impact scores for each fully connect neuron and a domain guided dropout based on the impact scores, the learned generic embedding's yielded competitive Re-ID accuracy. Su *et al.* [44] proposed a semi-supervised attribute learning framework which progressively boosted the accuracy of attributes by only using a limited number of labelled data. This framework involves a three-stages training. A deep CNN (dCNN) is first trained on an independent dataset labelled with attributes. Then it is fine-tuned on another dataset only labelled with person IDs by using our defined triplet loss. Finally, the updated dCNN predicts attribute labels for the target dataset, which is combined with the independent dataset for the final round of fine-tuning. Liao *et al.* [45] proposed a new sampling strategy in the training process and improved the triplet loss function with two weight parameters. This method reduces the computation cost caused by large samples. It also contributes to learn multi-modal features under cross-data condition.

The advantages of the CNN-based methods are they can effectively extract the abstract and discriminable features between different pedestrians. However, CNN-based methods lack the structure to extract temporal features, Re-ID rates in the surveillance video become lower. It can be seen from Table 1 that the rates of local-based methods are relatively low. The reason is that the local features are too hard to be used as criteria for identifying IDs. The characteristics of pedestrians are described more detailed in the multi-scale network models. The advantage of multi-model network architecture lies in that the generalisation performance is improved by increasing the training datasets sizes and sorts. Multi-scale network structure works better in the datasets. However, considering the complexity of the background in practical applications, multi-model network structure is worth of attracting attentions.

3.2 Person Re-ID based on CNN and RNN

RNN [46] is used to describe the sequence behaviour of a dynamic system. In this way, the state information is circulated in the network and the spatial information is extracted as well. Person videos contain a large number of person appearance features and movement information, including attitudes, angles, backgrounds, gaits and so on. The information can help to establish a better model to express person characteristics. CNNs can establish the mapping relationships among the data, but cannot analyse the relationships within the time signals and cannot process the temporal information from person videos in the monitoring equipment. In order to achieve good effects, the authors of [3, 47–51] proposed the person Re-ID methods based on CNN and RNN. The algorithm mainly extracts features of each person image in a set of sequences by CNN and sends them to RNN to extract the temporal features. The basic flow is shown in Fig. 8.

3.2.1 Multi-frame network structure: Compared with single-frame, multi-frame matching has more desirable results. Mclaughlin *et al.* [47] used multi-frame matching to solve person video matching problems. In order to reduce the computation cost and overfitting phenomenon, Mclaughlin *et al.* used the pooling operation to aggregate the frame-level features into a high generalisation global vector to reduce the feature dimension and the computation cost. Mclaughlin *et al.* proposed a Siamese network structure that combined CNN and RNN. CNN is used to extract the features of multi-frame data and optical flow data. Then passing the outputs chronologically through the RNN can get the person features in both spatial and temporal dimensions. Temporal pooling reduces the feature dimensions and using an identification loss can speed up the training process. This algorithm can deal with any length of video sequences and build an effective model to describe the spatial and temporal features. On the basis of this algorithm, Wu *et al.* [48] used the GRU instead of RNN in the network structure to process arbitrary length of video sequences and got the person spatial and temporal characteristics effectively.

3.2.2 Attention-based network structure: Matching methods based on multi-frame increases the network complexity and computational cost. In order to solve these problems, scholars have put forward to a variety of the attention models. Based on the structure of CNN + RNN, key features among multi-frame images are extracted by these attention models, and more discriminative images and image contents are selected from original images to achieve person Re-ID. In the meantime, the weights of both redundant and irrelevant features decrease, respectively. Liu *et al.* [49] put forward to the comparative attention network (CAN). The

Table 1 Comparison of the performances of different CNN-based methods

Method		PRID-2011	iLIDS-VID	VIPeR	Market-1501	CUHK03	CUHK01
local-based	TCP [28]	22.0	60.4	47.8	—	—	53.7
	S-CNN [31]	—	—	37.8	76.04	68.1	—
multi-scale	HP-Net [34]	—	—	—	76.9	—	—
	DC features [7]	—	—	—	86.79	74.21	—
	PIE [37]	—	—	27.44	78.65	62.60	
	Spindle Net [41]	67.0	66.3	53.8	76.9	88.5	79.9
	PDC [40]	—	—	51.27	84.14	88.70	
	HA-CNN [35]	—	—	—	91.2	44.4	—
multi-model	FT-JSTL + DGD [43]	64.0	64.6	38.6	—	75.3	66.6
	SSDAL [44]	20.1	—	37.9	—	—	—

whole CAN structure consists of two parts, global discriminant feature learning and local attention comparison learning. During the training phase, the CNN with shared weights is used to learn the global features from triple images, and then is sent to long short-term memory (LSTM) to obtain discriminative local visual attention features H, H+ and H− by comparing the positive and negative sample pairs. Finally, using the triple loss function to reduce the distance within same classes, as well as, to increase the distance between different classes. They calculated the distance of each pair of person and sorted them during the test phase. The architecture of the proposed CAN is shown in Fig. 9. However, this kind of method needs to browse the images several times in the query process, which can result in a large amount of computation costs. In the meaning time, the single-frame matching model has a lower generalisation. On the basis of network structure of CNN and RNN, Xu *et al.* [50] proposed Spatial and Temporal Attention Pooling Network, which contains the attentive spatial pooling structure and the attentive temporal pooling structure. The attentive spatial pooling structure includes four scales of the pyramid pooling layer. The attentive temporal pooling structure calculates an attention metric matrix which is a pair of pedestrians' output matrices through parallel RNNs and then processed by a maximum pooling. This algorithm not only can extract the interest regions of each frame, but also can find the key frames in a sequence and reduce the computation cost of neural network. Liu *et al.* [3] proposed an end-to-end algorithm called Accumulative Motion Context (AMOC) to accumulate motion context information. The algorithm mainly integrates person appearance features and motion context information through two sub-networks (motion network and spatial network). Person's valid information and dynamic points among adjacent frames are extracted by RNN. Finally, person matching and person Re-ID are achieved by introducing classification loss and verification loss. The motion network is composed of convolution and deconvolution layers, and the spatial network consists of two same convolutional structures used for learning the spatial and temporal features from the original video. The architecture is shown in Fig. 10. The advantage of this algorithm is that each pair of consecutive frames are processed by the Simeses network, and the deconvolution layers in the motion network can learn advanced feature expression. At the same time, the generalisation performance is outstanding under the unfavourable environment. Dai *et al.* [51] proposed a learning framework based on person video sequences, which can not only find and pay attention to the adequate temporal information from videos, but also solve the poor alignment of spatial information in the moving process. The framework includes temporal residual learning (TRL) module and a spatio-temporal transformer (ST^2N) module. TRL uses the bi-directional LSTM model and the residual method, which enable the network to describe a moving person and make full use of the

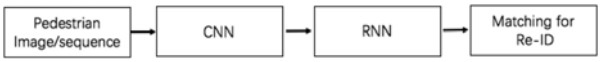

Fig. 8 *Re-ID system based on CNN and RNN diagram*

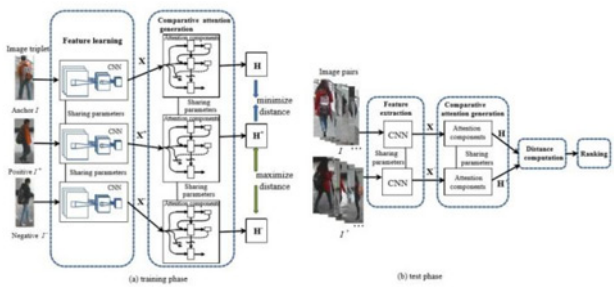

Fig. 9 *CAN [49]*

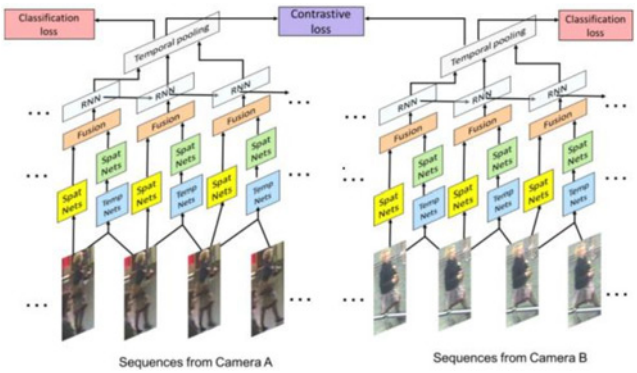

Fig. 10 *AMOC network [3]*

complementary information for feature extraction. ST^2N uses high-level semantic information and adjacent frames to train a model with smaller parameters, thereby solving the poor alignment problem under significant appearance changes.

The advantages of the algorithms based on CNN and RNN are that it can deal with the spatial relationship from pedestrian image contents through CNN and use RNN to process multi-frame image information to extract more accurate pedestrian characteristics. It can be seen from Table 2 that the recognition rates of attention-based models are higher in the different datasets. Because they can adaptively select the key parts from multi-frame images to describe pedestrian characteristics.

3.3 GAN-based methods

GAN [52] is an emerging technique for both semi-supervised and unsupervised learning. Goodfellow achieves this through deriving back propagation signals through a competitive process involving a pair of networks. They can be characterised by training a pair of networks in competing with each other. A common analogy, apt for visual data, is to think of one network as an art forger, and the other as an art expert. The forger, known in the GAN literature as the generator, *G*, creates forgeries, with the aim of making realistic images. The expert, known as the discriminator, *D*, receives both forgeries and real (authentic) images, and aims to tell them apart. Both are trained simultaneously. GAN can be used in a variety of applications, including image synthesis, semantic image editing, style transfer, image super-resolution, classification and person Re-ID. The GAN-based person Re-ID flow is shown in Fig. 11.

At present, there have been many papers that adopt GAN to solve the problems of person Re-ID. The existing datasets have low diversities and small scales, which leads to poor generalisation performance on the trained models. To solve this problem, Zheng *et al.* [53] used the GAN to generate new pedestrian images with new labels by semi-supervised learning. They marked the generated images with the LSRO label distribution. On Market-1501, they achieved rank-1 accuracy = 78.06%, mAP = 56.23%, and arrived at rank-1 accuracy = 73.1%, mAP = 77.4% on CUHK03, which are all very competitive. The innovate idea of this method is to increase the generalisation ability by generating new samples with LSRO label. Yet new samples are too blurry to meet artificial benchmarks and cannot increase the data size

Table 2 Comparison of the performances of different CNN and RNN-based methods

Method		PRID-2011	iLIDS-VID	MARS
multi-frame	CNN + RNN [47]	70	58	40
	deep RCN [48]	49.8	42.6	
attention-based	AMOC [3]	83.7	68.7	
	ASPTN [50]	77	62	44
	TRL + ST^2N [51]	87.8	57.7	79.3

Fig. 11 *Re-ID system based on GAN diagram*

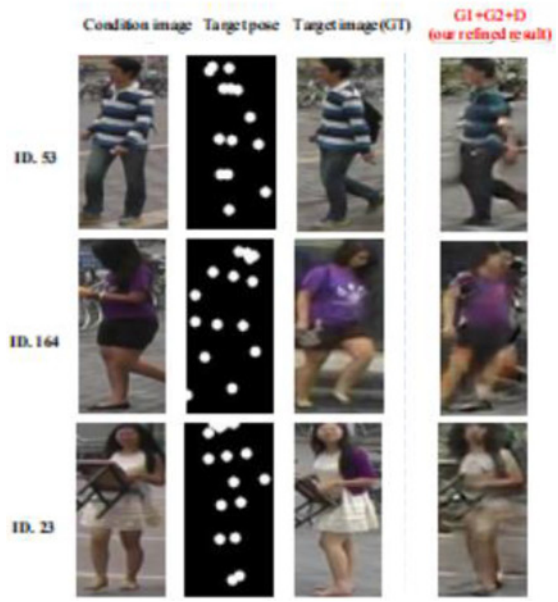

Fig. 12 *Results on market-1501 by PG^2 [54]*

directly. However, they are the first people who applied GAN to person Re-ID. Ma *et al.* [54] proposed Pose Guided Person Image generation network (PG2). It can generate a clear pedestrian image (Fig. 12) with a specified posture on the base of pose guidance. The algorithm is divided into two stages. The first stage generates the initial image which has the target pose, as well as the person global structure. In the second stage, the initial image and the input image are combined and sent to another U-net-like. Then the initial image is refined by the confrontation training. Ma *et al.* proposed to use the pose mask to calculate the loss function of the first stage generation, so it can alleviate background influences on the generated image. They used the U-net [55] network with removing the full connection layer. By this way, the information of the input image can be retained as much as possible, which reduces features loss during information transfer.

Yin *et al.* [56] used GAN to study person Re-ID from the perspective of semantic attribute recognition and improved the cross-modal matching recognition rate. The model optimises the attribute similarity C^A of the input image through semantic consistency constraints and adversary loss, and optimises the attribute expression of the input image by image concept extraction loss. The model can finally obtain the optimal semantic description of the input image, which is used as a criterion for person Re-ID (Fig. 13).

The number of GAN-based methods is not enough. However, the novel ideas are ways to solve the person Re-ID problems. The PG^2 network can generate pedestrian images with arbitrary poses, which provides the possibility to make datasets larger. The recognition rates achieved by the GAN-attribute recognition network also exceed the most existing algorithms.

3.4 Hybrid methods

The hybrid method means deep models combined with the traditional methods, such as Local Binary Pattern (LBP), LOMO and so on. Features can be automatically extracted by the deep network, but the too deep structure is not conducive for training parameters, as well as spatial information is constantly diluted. So the papers [3, 7, 11, 57, 58] introduced the traditional methods to

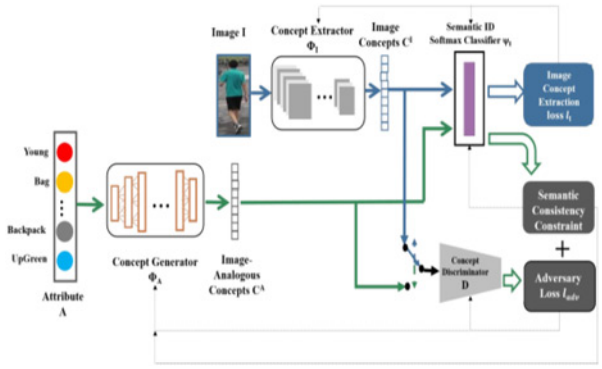

Fig. 13 *Structure network proposed by Yin et al. [56]*

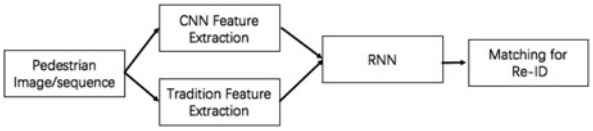

Fig. 14 *Re-ID system based on the hybrid diagram*

improve model performance and had achieved good results. The basic flowchart of the hybrid method is shown in Fig. 14.

Li *et al.* [7] combined CNN with LOMO, resulting in a rank-1 accuracy of 22.6% on the PRID dataset, which is an increase of 2.5% over the deep-learning-only method. Liu *et al.* [3] proposed the AMOC algorithm on the basis of the EpicFlow [59] method, with rank-1 reaching 68.7% on the iLIDS-VID [6] dataset. Yan *et al.* [57] put forward to the recurrent feature aggregation network model. Firstly, the colour and LBP features at t moment are extracted to a feature cascade. Then the person characteristics of a video sequence are learned by LSTM modules. At last, error calculation and gradient back propagation are achieved by the softmax layer. The model outperformed the recognition rates of LBP colour features and non-deep learning algorithms. The detailed structure is shown in Fig. 15.

In order to solve the correlation problem between the feature vectors after the full connected layer, Sun *et al.* [58] proposed a CNN-based model called SVDnet. They used the singular vector decomposition to optimise the deep feature representation and the relaxation iteration strategy to integrate orthogonal constraints in network training and to decouple feature vectors. This model achieved the rank-1 rate of 80.5% on Market-1501 [8]. Bak and Carr [10] used One-Shot [60] method to measure the similarity between person images from texture and colour feature. This method uses CNN to extract the features of person grayscale images from multiple datasets, which reduces the correlation between texture features and colour features. Learning the person images' the local similarity measure matrix by the colour card table can effectively solve the colour difference problem from multiple cameras. The advantage of this method is that only a few

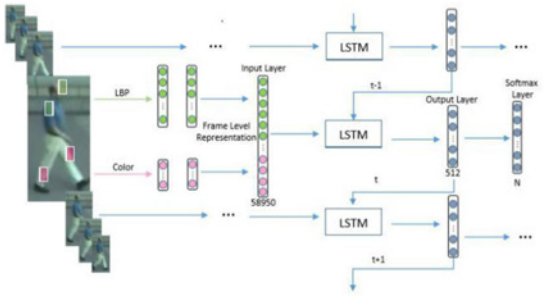

Fig. 15 *Detailed structure of the framework based on LSTM [57]*

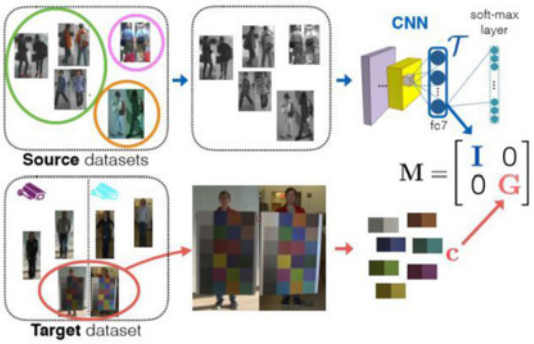

Fig. 16 *One-shot metric learning algorithm structure [10]*

training samples are needed, which avoids the dependence of data size. The specific operation is shown in Fig. 16.

Deep learning means learning weight and bias parameters in neural networks through a large number of training samples. However, too many parameters can easily lead to overfitting and reduce the Re-ID accuracy. Moreover, the pedestrian characteristics extracted by the neural networks have no obvious realistic meanings. Traditional feature-based methods are mostly on the basis of human's understanding on images and can describe image contents quickly and accurately. The advantages of the hybrid methods are that it can speed up the learning process of the neural networks and ensure that the network models learn more effective pedestrian features. The introduction of other traditional pattern recognition algorithms can also compress the size of the neural networks and reduce the number of parameters, as well as overfitting. It can be seen from Table 3 that the recognition rates of the hybrid methods are higher than those of the deep-only methods.

3.5 Performance comparison

The common frameworks for person Re-ID include person feature extraction and matching. The combination of different feature extraction methods and matching strategies will produce different effects. However, if the extracted features are not discriminative, the final matching result will be affected. By constructing multiply layers network, deep learning [2] is able to learn the characteristics from Big Data. It can learn an abstract representation from the data and find the implicit information. Compared with the traditional methods, deep learning uses its own deep structure to learn features automatically from a large amount of data. At the same time, the combination of traditional algorithms and deep learning can solve the person Re-ID problems more effectively. At present, the highest recognition rates reached by the person Re-ID algorithms in the mainstream datasets are shown in Table 4. It can be clearly seen from the table that the highest recognition rates achieved by the deep methods in the mainstream datasets exceed the highest recognition rates based on the traditional methods.

4 Difficulties and future

4.1 Difficulties

Person Re-ID based on deep learning is still in the development stage and cannot be completely put into practical application. At present, the methods introduced with deep learning are mostly on the basis of improvements on the existing basic network models. According to person image characteristics, establishing an innovate deep network model becomes a main research direction. Although these approaches have achieved good results, it is wasted to constantly adjust the parameters in the training process, and it is easy to occur overfitings when network structure becomes deep. Moreover, reducing the computation costs and saving time as far as possible are the keys to practical. As the number of training samples affects the deep learning models performance, only establishing a large-scale standard dataset can ensure that the trained models still have a good generalisation ability under the complex environment. At the same time, as a multi-camera and cross-view recognition application, person Re-ID will be applied in reality occasions. However, the existing algorithms are difficult to ensure the recognition accuracy and the rapidity at the same time.

4.2 Future

(i) *End-to-end:* Existing person Re-ID databases are all established in the ideal conditions and person samples are detected from videos by hand or detection algorithms automatically. However, facing the massive video data in reality, constructing an end-to-end method including pedestrian detection and person Re-ID is one of the research trends. At present, there are some methods and research on this area. Zheng *et al.* [61] proposed the mechanisms for pedestrian detection to help improve overall Re-ID accuracy, and assessed the effectiveness of different detectors for Re-ID. They established an end-to-end person Re-ID network structure with raw videos as the inputs. The fine-tuning strategy is used to train pedestrian detection model and classification model.
(ii) *Network structure:* The network structures that effectively extract person characteristics still have research value. Person Re-ID needs to recognise the same person ID from different angles, clothes, poses and complex backgrounds. Attention-based network structure can choose significant features from images and video sequences, as well as

Table 3 Comparison of the performances of different hybrid methods

	Method	PRID-2011	iLIDS-VID	MARS	VIPeR	Market-1501	CUHK03
only deep	ResNet-50 [58]	—	—	—	—	73.8	66.2
	CAN (VGG-16) [49]	—	—	—	47.2	72.1	77.6
hybrid	end-to-end AMOC + EpicFlow [3]	83.7	68.7	68.3	—	—	—
	SVDNet + ResNet-50 [58]	—	—	—	—	82.3	81.8
	CAN(VGG-16) + LOMO [49]	—	—	—	54.1	—	—
	one-shot [10]	41.4	51.2	—	34.3	—	—

Table 4 Comparison on traditional person Re-ID algorithms and deep learning Re-ID algorithms

Dataset	Deep method	Rank 1	Rank 5	mAP	Traditional method	Rank 1	Rank 5	mAP
PRID-2011	TRL + ST^{2}N [51]	87.8	97.4		TDL [4]	57.74	80.00	—
iLIDS-VID	AMOC [3]	68.7	94.3	—	TDL [4]	56.33	87.60	—
MARS	DC features [7]	83.03	93.69	66.43	k-reciprocal encoding [22]	73.94	—	68.45
VIPeR	Spindle Net [41]	62.1	83.4	—	CSBT [23]	36.6	66.2	—
Market-1501	HA-CNN [35]	91.2	—	75.7	k-reciprocal encoding [22]	77.11	—	63.63
CUHK03	HP-Net [33]	91.8	98.4	—	k-reciprocal encoding [22]	69.90	—	72.45
CUHK01	end-to-end [49]	87.2	98.2	—	CSBT [23]	51.2	76.3	—

remove redundant information. Multi-model network structure learns a variety of datasets to increase the generalisation performance. The multi-scale network structure can fully extract feature expressions. The above structures all have their corresponding advantages, but their scales have gradually increased, which add difficulties in training and adjustment process, as well as increase the cost of R&D work. It is of great value to improve the Re-ID accuracy on the basis of shrinking the network structure.

(iii) *GAN:* As GAN widely used in the fields of style transfer, image generation, detection and classification, there have been some scholars using GAN to research person Re-ID. First, the number of the training samples is insufficient, as well as lacking of diversity. The performances of deep learning models using CNN and RNN also depend on the quality and quantity of samples. At the same time, the labour cost of establishing a standard dataset is large. In order to solve these problems, Ma *et al.* [54] have proposed a PG2 network structure to convert a pedestrian image into another one with an arbitrary pose, but the complex pose transfer results require to be improved. Moreover, it is possible to generate pedestrian images with any kinds of characters in the future. Yin *et al.* [56] and Zheng *et al.* [53] proposed Re-ID frameworks with GAN. Zheng *et al.* [53] proposed to use DCGAN to generate images with new labels for training. It can improve the generalisation with non-standard pedestrian images. Yin *et al.* [56] used GAN to obtain the optimal semantic attribute vector from the image and used the semantic attribute vector for discriminating identity. As an emerging technique for both semi-supervised and unsupervised learning, GAN gives another innovative idea for person Re-ID. Therefore, adopting GAN may achieve far more results than the CNN and RNN structures.

(iv) *Rapidity and accuracy:* Most of the existing methods have limitation, which is hard to guarantee the rapidity and accuracy in the practical application. Lin *et al.* [62] proposed a more practical matching algorithm called CADL, which calculates the entire monitoring network. It uses CNN to extract person characteristics from multiple cameras and calculates the cosine similarity of the person feature maps between different cameras. Finally, the similarity matrix is used for gradient descent and back propagation. The algorithm's ranks-1 reached to 80.85% on the Maket-1501 [8] dataset. This method obtains the global optimal solution and can balance the recognition performance under different cameras, which are beneficial for the application of person Re-ID. The algorithms based on deep learning take both the rapidity and accuracy into account will have great research value.

5 Conclusion

This paper first introduces the traditional methods of person Re-ID, and analysis its advantages and disadvantages. It mainly focuses on deep learning methods in recent years. Deep learning uses its own deep structure to learn features automatically from a large amount of data, avoiding the shortcomings of traditional manual feature extraction methods. It also extracts the advanced semantic features and temporal features. Deep learning is more robust than most traditional methods; the combination of deep learning and traditional methods is very significant. In terms of person feature extraction, some algorithms have achieved remarkable results. Although deep learning has achieved good results in the field of image classification and speech recognition, the application of person Re-ID is not yet mature. As many problems exist, there are still some valuable research directions in the future. We need to constantly focus on how to apply deep learning better to make person Re-ID applied to the actual situation.

6 Acknowledgment

This work was supported by the Natural Science Foundation of China No. 61703119, 61573114, Natural Science Fund of Heilongjiang Province of China No. QC2017070 and Fundamental Research Funds for the Central Universities of China No. HEUCFM180405.

6 References

[1] Bedagkar-Gala, A., Shah, S. K.: 'A survey of approaches and trends in person re-identification', *Image Vis. Comput.*, 2014, **32**, (4), pp. 270–286
[2] Hinton, G.E., Salakhutdinov, R.R.: 'Reducing the dimensionality of data with neural networks', *Science*, 2006, **313**, (5786), pp. 504–507
[3] Liu, H., Jie, Z., Jayashree, K., *et al.*: 'Video-based person Re-identification with accumulative motion context', *IEEE Trans. Circuits Syst. Video Technol.*, 2017, **PP**, (99), pp. 1–1
[4] You, J., Wu, A., Li, X., *et al.*: 'Top-push video-based person re-identification'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1345–1353
[5] Liu, K., Ma, B., Zhang, W., *et al.*: 'A spatio-temporal appearance representation for video-based pedestrian Re-identification'. IEEE Int. Conf. on Computer Vision, Santiago, Chile, 2015, pp. 3810–3818
[6] Zheng, W. S., Gong, S., Xiang, T.: 'Associating groups of people'. British Machine Vision Conf., BMVC 2009, Proc. DBLP, 2009, London, UK, 7–10 September 2009
[7] Li, D., Chen, X., Zhang, Z., *et al.*: 'Learning deep context-aware features over body and latent parts for person re-identification'. IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii, USA, 2017, pp. 384–393
[8] Zheng, L., Shen, L., Tian, L., *et al.*: 'Scalable person re-identification: a benchmark'. IEEE Int. Conf. on Computer Vision, Santiago, Chile, 2016, pp. 1116–1124
[9] Jose, C., Fleuret, F.: 'Scalable metric learning via weighted approximate rank component analysis'. European Conf. on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 875–890
[10] Bak, S., Carr, P.: 'One-shot metric learning for person re-identification'. IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii, USA, 2017, pp. 1571–1580
[11] Chakraborty, A., Das, A., Roychowdhury, A.: 'Network consistent data association', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (9), pp. 1859–1871
[12] Liao, S., Hu, Y., Zhu, X., *et al.*: 'Person re-identification by local maximal occurrence representation and metric learning', Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 2015, pp. 2197–2206
[13] Zhao, R., Ouyang, W., Wang, X.: 'Unsupervised salience learning for person re-identification'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 3586–3593
[14] Lowe, D. G.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis.*, 2004, **60**, (2), pp. 91–110
[15] Paisitkriangkrai, S., Shen, C., van den Hengel, A.: 'Learning to rank in person re-identification with metric ensembles'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 2015, pp. 1846–1855
[16] Chen, L., Chen, H., Li, S., *et al.*: 'Person Re-identification by color distribution fields', *J. Chi. Comput. Syst.*, 2017, **38**, (6), pp. 1404–1408 (in Chinese)
[17] Yang, M., Wan, W., Hou, L., *et al.*: 'Person re-identification using human salience based on multi-feature fusion'. Int. Conf. on Smart and Sustainable City and Big Data, Shanghai, China, 2016, pp. 1–5
[18] Geng, Y., Hu, H.-M., Zeng, G.: 'A person re-identification algorithm by exploiting region-based feature salience', *J. Vis. Commun. Image Represent.*, 2015, **29**, (C), pp. 89–102
[19] Layne, R., Hospedales, T.M., Gong, S., *et al.*: 'Person re-identification by attributes', BMVC, London, UK, 2012, vol. 2, (3), p. 8
[20] Weinberger, K.Q., Saul, L. K.: 'Distance metric learning for large margin nearest neighbor classification'. JMLR.org, 2009, 10, (1), pp. 207–244
[21] Yu, H.X., Wu, A., Zheng, W. S.: 'Cross-view asymmetric metric learning for unsupervised person re-identification', arXiv preprint arXiv:1708.08062, 2017, pp. 994–1002
[22] Zhong, Z., Zheng, L., Cao, D., *et al.*: 'Re-ranking person re-identification with k-reciprocal encoding', The IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii, USA, 2017, pp. 318–1327
[23] Chen, J., Wang, Y., Qin, J., *et al.*: 'Fast person re-identification via cross-camera semantic binary transformation'. IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii, USA, 2017, pp. 3873–3882
[24] Li, W., Zhao, R., Xiao, T., *et al.*: 'DeepReID: deep filter pairing neural network for person re-identification'. *Comput. Vis. Pattern Recognit.*, 2014, pp. 152–159
[25] Yi, D., Lei, Z., Liao, S., *et al.*: 'Deep metric learning for person re-identification'. Int. Conf. on Pattern Recognition, Columbus, OH, USA, 2014, pp. 34–39
[26] Lecun, Y., Bottou, L., Bengio, Y., *et al.*: 'Gradient-based learning applied to document recognition'. Proc. of the IEEE, Oakland, 1998, pp. 2278–2324
[27] Li, W., Zhao, R., Xiao, T., *et al.*: 'Deepreid: deep filter pairing neural network for person re-identification'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 152–159
[28] Cheng, D., Gong, Y., Zhou, S., *et al.*: 'Person re-identification by multi-channel parts-based CNN with improved triplet loss function'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1335–1344
[29] Li, S., Chen, L.: 'Person re-identification based on locally deep matching', *Appl. Res. Comput.*, 2017, **34**, (4), pp. 1235–1238. (In Chinese)
[30] Felzenszwalb, P.F., Girshick, R.B., Mcallester, D., *et al.*: 'Object detection with discriminatively trained part-based models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **47**, (2), pp. 6–7
[31] Varior, R.R., Haloi, M., Wang, G.: 'Gated Siamese convolutional neural network architecture for human re-identification'. European Conf. on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 791–808

[32] Ahmed, E., Jones, M., Marks, T. K.: 'An improved deep learning architecture for person re-identification'. The IEEE Conf. on Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 2015, pp. 3908–3916
[33] Qian, X., Fu, Y., Jiang, Y.G., *et al*.: 'Multi-scale deep learning architectures for person re-identification'. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 5409–5418
[34] Liu, X., Zhao, H., Tian, M., *et al*.: 'Hydraplus-net: attentive deep features for pedestrian analysis'. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 350–359
[35] Li, W., Zhu, X., Gong, S.: 'Harmonious attention network for person re-identification', arXiv:1802.08122v1, 2018
[36] Jaderberg, M., Simonyan, K., Zisserman, A.: 'Spatial transformer networks'. *Adv. Neural. Inf. Process. Syst.*, Istanbul, Turkey, 2015, pp. 2017–2025
[37] Zheng, L., Huang, Y., Lu, H., *et al*.: 'Pose invariant embedding for deep person re-identification', arXiv preprint arXiv:1701.07732, 2017
[38] Dong, S.C., Cristani, M., Stoppa, M., *et al*.: 'Custom pictorial structures for re-identification'. British Machine Vision Conf., Dundee, UK, 2011, pp. 68.1–68.11
[39] He, K., Zhang, X., Ren, S., *et al*.: 'Deep residual learning for image recognition'. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778
[40] Su, C., Li, J., Zhang, S., *et al*.: 'Pose-driven deep convolutional model for person re-identification'. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3980–3989
[41] Zhao, H., Tian, M., Sun, S., *et al*.: 'Spindle net: person re-identification with human body region guided feature decomposition and fusion'. Computer Vision and Pattern Recognition, Hawaii, USA, 2017, pp. 1077–1085
[42] Ren, S., He, K., Girshick, R., *et al*.: 'Faster R-CNN: towards real-time object detection with region proposal networks'. Int. Conf. on Neural Information Processing Systems, Palais des Congrès de Montréal, 2015, pp. 91–99
[43] Xiao, T., Li, H., Ouyang, W., *et al*.: 'Learning deep feature representations with domain guided dropout for person re-identification'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1249–1258
[44] Su, C., Zhang, S., Xing, J., *et al*.: 'Deep attributes driven multi-camera person re-identification'. European Conf. on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 475–491
[45] Liao, W., Yang, M.Y., Zhan, N., *et al*.: 'Triplet-based deep similarity learning for person re-identification', 2018
[46] Kuhn, G., Watrous, R.L., Ladendorf, B.: 'Connected recognition with a recurrent network', *Speech Commun.*, 1990, **9**, (1), pp. 41–48
[47] Mclaughlin, N., Rincon, J.M.D., Miller, P.: 'Recurrent convolutional network for video-based person re-identification'. IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1325–1334
[48] Wu, L., Shen, C., Hengel, A.: 'Deep recurrent convolutional networks for video-based person re-identification: an end-to-end approach', arXiv preprint arXiv:1606.01609, 2016
[49] Liu, H., Feng, J., Qi, M., *et al*.: 'End-to-end comparative attention networks for person re-identification', *IEEE Trans. Image Process.*, 2017, **PP**, (99), pp. 1–1
[50] Xu, S., Cheng, Y., Gu, K., *et al*.: 'Jointly attentive spatial-temporal pooling networks for video-based person re-identification'. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 4743–4752
[51] Dai, J., Zhang, P., Lu, H., *et al*.: 'Video person re-identification by temporal residual learning', arXiv:1802.07918, 2018
[52] Goodfellow, I.J., Pougetabadie, J., Mirza, M., *et al*.: 'Generative adversarial networks', *Adv. Neural. Inf. Process. Syst.*, 2014, **3**, pp. 2672–2680
[53] Zheng, Z., Zheng, L., Yang, Y.: 'Unlabeled samples generated by GAN improve the person re-identification baseline in vitro'. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3774–3782
[54] Ma, L., Jia, X., Sun, Q., *et al*.: 'Pose guided person image generation', arXiv:1705.09368, 2017
[55] Quan, T.M., Hilderbrand, D.G.C., Jeong, W. K.: 'Fusionnet: a deep fully residual convolutional neural network for image segmentation in connectomics', arXiv:1612.05360, 2016
[56] Yin, Z., Zheng, W.S., Wu, A., *et al*.: 'Adversarial attribute-image person re-identification', arXiv:1712.01493, 2017
[57] Yan, Y., Ni, B., Song, Z., *et al*.: 'Person re-identification via recurrent feature aggregation'. European Conf. on Computer Vision, 2016, pp. 701–716
[58] Sun, Y., Zheng, L., Deng, W., *et al*.: 'SVDNet for pedestrian retrieval'. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3820–3828
[59] Revaud, J., Weinzaepfel, P., Harchaoui, Z., *et al*.: 'Epicflow: edge-preserving interpolation of correspondences for optical flow'. Computer Vision and Pattern Recognition, Boston, Massachusetts, USA, 2015, pp. 1164–1172
[60] Fei-Fei, L., Fergus, R., Perona, P.: 'One-shot learning of object categories', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (4), pp. 594–611
[61] Zheng, L., Zhang, H., Sun, S., *et al*.: 'Person re-identification in the wild'. IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017, pp. 3346–3355
[62] Lin, J., Liang, L., Ren, L., *et al*.: 'Consistent-aware deep learning for person re-identification in a camera network'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii, USA, 2017, pp. 5771–5780