Overview of deep learning based pedestrian attribute recognition and re-identification

Duidi Wu, Haiqing Huang *, Qianyou Zhao, Shuo Zhang, Jin Qi, Jie Hu

Institute of Knowledge Based Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

1. Introduction

Pedestrian re-identification can be considered as a research area for image retrieval, which uses computer vision to judge whether the target pedestrian is present in specified image or in consecutive frames. Given a monitored pedestrian image, the purpose is to retrieve the pedestrian image across devices. Pedestrian re-recognition technology can effectively use multiple camera views, often complementing pedestrian detection and pedestrian tracking, which is of great importance in the fields of intelligent surveillance and security.

General pedestrian re-identification can be divided into four steps, as shown in Fig. 1, including pedestrian detection, feature extraction, similarity measurement and image retrieval respectively. The input of pedestrian re-recognition is a specific pedestrian image called probe. Feature extraction is to learn the diverse features of pedestrians on different cameras. Metric learning is to map the features to a new space, so that people with similar features are clustered together. Image retrieval is to sort and reranking pictures according to the similarity distance between features.

The traditional algorithm of pedestrian ReID is mainly based on manual design features combined with distance measurement. The research of feature extraction mainly focuses on obtaining robust and discriminative low-level visual features, including color features or histograms in RGB and HSV color spaces. Texture features (Gabor) [3] Shape features (HOG features) [4] and key points (SIFT) [5] and so on. Metric learning finds feature mappings in different spaces, so that the feature vector of the same kind of sample is closer than the feature vector of the non-same kind of sample. Commonly used metric learning methods include Mahalanobis distance [1], local adaptive decision function [6], and saliency weighted metric learning [2]. Due to the pedestrian re-recognition problem, the scene (background, posture, etc.) changes are complex and other challenges.

The above-mentioned traditional low-level visual feature extraction methods are difficult to obtain effective and invariant features, which directly affect the accuracy and effectiveness of metric learning methods. In this paper, we mainly explore pedestrian ReID based on deep learning, which use more advanced algorithms to extract discriminative features. The preprint [7] is available at SSRN.

* Corresponding author.
E-mail addresses: achelils@sjtu.edu.cn (H. Huang).

https://doi.org/10.1016/j.heliyon.2022.e12086
Received 1 April 2022; Received in revised form 6 July 2022; Accepted 25 November 2022

2405-8440/© 2022 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
2. Association between APR and ReID

2.1. Attribute-Assisted re-identification

From the perspective of the composition of the ReID system, the realization of ReID requires feature extraction and then similarity measurement, attribute recognition also needs to extract pedestrians features. Attribute recognition network usually uses CNN to extract the features. Generally, the output of the model is consistent with the number of attributes needed to be recognized. The output of pedestrian ReID is generally not directly equal to the extracted features. Most methods use attribute to assist re-recognition [8, 9, 10, 11, 12] as shown in Fig. 2. Yin et al. [13] firstly presented extensive investigation on the attribute-image ReID problem under an adversarial framework.

One application of ReID is finding a suspect of interest in a criminal investigation. Witnesses often indicate personal traits of the suspect as seen at the moment when the crime was committed, including hair color, clothing type, etc. Based on that description, the police manually scan the entire video archive looking for a person with similar characteristics. Since people are less likely to change clothes or shave when walking in two different places in a short period of time, so attributes can be used for tracking and communication between two related cameras to achieve re-identification.

Human attributes usually stand as robust visual properties when condition changes and have been investigated in many works. This part analyzes several methods that leverage features from human attributes for person ReID. The algorithms are summarized in Fig. 4.

2.1.1. Attribute attention mechanism

Attention models have been used in fine-grained recognition to crop the key parts. In Comparative Attention Network (CAN) [14], attention mechanism was used in person re-identification by integrating a recurrent attention with the Siamese model. Attribute-Aware Attention Model (AAM) [8] explored attention mechanism to select better attribute features. To generate attentions for global category features fusion and local attribute features fusion simultaneously, it consists of two components: attribute-guided attention module and category-guided attention module. The final feature is the addition of the weighted category feature and attribute feature $f_{\text{final}} = f_{\text{region}} + f_{\text{attr}}$.

2.1.2. Attribute feature fusion

Layne et al. [9] proposed Attribute Interpreted ReID (AIR) which assigned weights to semantic attributes to serve as a representation of pedestrian features. The model learned an attribute-centric, parts-based feature representation, by selecting those attributes that are most effective for re-identification, the attribute-level information and standard low-level features are fused. Attributes-aided Part Detection and Refinement model (APDR) [11] designed an effective attribute fusion network, extracted local attribute features to handle the body part misalignment problem, which is another major challenge for ReID. Experiments show that the learned local features, along with holistic image-level feature, can further improve the accuracy on ReID task. In Fig. 3, Fariborz et al. [15] used a tensor to non-linearly fuse identity and attribute features, and then forced the parameters of the tensor in the loss function to generate discriminative fused features for ReID. Lin et al. [16] proposed the ARM module to calculate the weighting of $M$ individual attributes, which effectively complements the global features for ID loss.

2.1.3. Attribute search / zero-shot ReID

Attributes provide a low-dimensional mid-level representation which variability in appearance. Although individual attributes vary in robustness and informativeness, they provide a strong cue for identity. Semantic attributes could potentially be used to constrain or permute a search for a particular person, for example by specifying invariance to
whether or not they have removed or added a hat. Given an attribute description in the form of a binary vector, a person can be found by NN matching against the attribute profiles $A_{x_i}$ of each person $i$ in the dataset [16]. In the short-term recognition, Parts-and-Attributes Search Framework (PASF) [17] distinguished people based on parsing the partial human body attributes, visual attributes include hair type (beards, mustaches, absence of facial hair), type of eye-wear (sunglasses, eyeglasses, absence of glasses), hair type (baldness, hair, wearing a hat), and clothing color. Face detection result was matched to its neighboring regions based on correlation, each person who leaves the scene is assigned a unique identifier (“track ID”). Lampert et al. [18] collected judgments on the “relative strength of association” between 85 attributes and 48 animal classes, showed that description of high-level attributes makes sense for person ReID on untrained images. Optimized Attribute Re-identification (OAR) [19] showed mid-level attributes can be an effective representation for ReID and zero-shot identification.

2.1.4. Attribute fine-tuning

Su et al. [12] proposed SSDAL algorithm (Semi-supervised Deep Attribute Learning) and further extended [20]. First, train a deep Convolutional Neural Network on an independent dataset $T = t_1, t_2, ..., t_M$, in which each sample is labeled with a binary attribute label. Then it was fine-tuned on the dataset with person ID labels as $U = u_{1}, u_{2}, ..., u_{N}$, using triplet loss, where $N$, $M$ represent sample size. The third stage fine-tuned the model using sigmoid cross entropy loss on the dataset $T$ & $U$. Experiments showed that fine-tuning the model trained on PAR dataset acquires more discriminative power for ReID task. Attribute-complementary re-id net (ACRN) [21] included automatically predicted attribute features into the training stage of the model. This allowed the CNN to focus on learning information for ReID which is complementary contained in the attribute predictions. Tetsu et al. [22] et al. improved the CNN features by conducting a fine-tuning on a pedestrian attribute dataset, proposed a loss function for classifying combination attributes to increase discriminative ability.

2.1.5. Attribute consistent embedding / space mapping

Attribute-Consistent Matching (ACM) [23] projected the input image into a lower dimensional subspace where the matches exhibit consistency to changing illumination. Su et al. [24] studied the tracklet-to-tracklet identification, proposed a novel discriminative model to exploit low-level features and map attributes to a discriminative latent prototypes space (LPSM). MTL-LORAE framework [25] used semantic attributes in addition to low-level features, and transforms the attributes in low rank space. Adaptive Semantic Margin Regularizer (ASMR) [26] conducted person search by finding images that are closest to the query in a co-embedding space of person images and categories. Wang et al. [27] introduced TJ-AIDL and IIA Space for achieving the knowledge fusion learning on attribute and ID labels.

2.1.6. Clothing attributes assisted

Li et al. [28] studied clothing attributes assisted person reidentification, presented latent support vector machine (c-LSVM)-based person ReID approach to describe the integration of low-level, mid-level and high-level attributes, specifically, it learns the relationship between body parts, clothing information and ID labels. Yang et al. [29] presented a discriminative latent model (DLM) for joint modeling of object classes and their visual attributes. Wu et al. [30] suggested the usefulness of clothing color in surveillance scenarios, and evaluated at the pixel level and frame level respectively. Dual Attribute-aware Ranking Network (DARN) [31] could retrieve clothing items that have the same or similar attributes as a given clothing image.

2.2. Differences between attribute recognition and ReID

2.2.1. Label annotation

Person ReID and attribute recognition usually use different datasets separately to train the network, which have different annotations for the dataset labels. Datasets with pedestrian attribute tags such as the Market-1501 [32] can be used universally. What needs to be paid attention to is that the specific usage may be different for the same dataset. Specifically, PAR focuses on the correspondence between images and tags, while pedestrian ReID focuses on personal identity information, in particular, divide the train and test based on ID.

2.2.2. Image feature

Many ReID algorithms combine local and global features such as PIE [33], PDC [34], GLAD [35], or merge local features such as Spindle Net [36]. These local features may be extracted through local blocks or according to methods such as posture key points, and not necessarily attribute features. Typically, the APR [16] network combined attribute loss and ID loss, which is also a typical example of Attribute-Assisted ReID.

2.2.3. Main focus of the model

Pedestrian re-identification is to find the image of the target pedestrian in the gallery data, so the model focuses on the similarity arrangement of the recognition results and the Average Precision of each query. The commonly used evaluation metrics mainly include CMC and mAP.

Attribute recognition aims to identify the attribute characteristics of pedestrian images as accurately as possible, so it focuses on the difference between the true label and the predicted value evaluate the model. Accuracy, precision, recall and F1 score are commonly used evaluation indicators.

In summary, Table 1, 2 shows the connection and difference between the two research fields of attribute recognition and pedestrian re-identification.

### Table 1. Comparisons of datasets and evaluations between PAR and ReID.

| Comparison | ReID | PAR |
|------------|------|-----|
| Datasets   | Market-1501 [32] | PETA [37] |
| Duke-MTMC  | 38    | PA-100K [39] |
| CUHK03     | 40    | Market-1501-Attribute [16] |
| Evaluation | CMC   | Precision |
|            | mAP   | Recall   |

### Table 2. Comparisons of labels and focus between PAR and ReID.

| Comparison | ReID | PAR |
|------------|------|-----|
| labels     | pedestrian ID arrangement of the recognition results | pedestrian attributes difference between true labels and predicted value |
| Focus      | attribute | attribute |
|            | label     | label    |
| top        | 1         | hair     |
| backpack   | 2         | up       |
| bag        | 1         | down     |
| handbag    | 1         | clothes  |
| boots      | 2         | hat      |
| shoes      | 1         | backpack |
| upblack    | 2         | bag      |
| downblue   | 2         | handbag  |
| age        | 2         |         |
| downwarded | 2         |         |

### Table 3. Sample of Pedestrian Image Attribute Annotation of DukeMTMC-Attribute and Market-1501-Attribute.

| DukeMTMC-Attribute | Market-1501-Attribute |
|--------------------|----------------------|
| attribute label    | attribute label      |
| top                | 1                    |
| backpack           | 2                    |
| bag                | 1                    |
| handbag            | 1                    |
| boots              | 2                    |
| shoes              | 1                    |
| upblack            | 2                    |
| downblue           | 2                    |
| age                | 2                    |
| downwarded         | 2                    |
or a sequence as shown in Table 5 and 6, respectively. The former category has Market-1501 [32], CUHK03 [40], DukeMTMC-reID [38], MSMT-17 [41], ViPeR [42], MARS [43], etc. Typical of the latter are iLIDS [44], Duke-Video [45], LPW [46], etc. Besides, there are other scene-based datasets, such as PRW [47]. Table 3 and Table 4 provide the examples and statistical summaries of these datasets.

As for the division of datasets, in the research of pedestrian re-identification, the datasets are usually obtained through manual annotation or detection algorithm. The pictures of pedestrians will be divided into training sets and validation sets, as well as query and gallery. After training, the model calculates the similarity for the extracted features in the query and the gallery separately. Generally, in the process of training and testing, the identity of the pedestrian is not repeatedly selected.

3.2. Attribute datasets

Compared with re-recognition datasets, attribute recognition focuses more on marking the appearance characteristics of pedestrians, such as gender, age, clothing, accessories, etc. Table 4 summarizes several commonly used datasets for attribute recognition. Fig. 5 concludes the annotation information of several APR datasets. Generally the datasets are labeled with Arabic numerals, PETA [37] and PA-100K [39] using 1 and 0, existing attributes are represented by 1, labels in Market-1501 Attribute [16] and DukeMTMC Attribute [16] use 1,2,3,4 to represent each type of attribute. Table 3 gives an example of annotation.

3.3. Metrics to measure ReID experiment results

To evaluate an attribute recognition system, Precision (P) and Recall (R) are two widely used measurements. Moreover, mAP and CMC are two indicators commonly used to measure re-identification and retrieval capabilities.

The attribute recognition task mainly focuses on the differences between the true labels and predicted values. Both Precision (P) and Recall (R) can reflect the exact numbers of items retrieved, but they are actually different. Specifically, Precision is the number of correct information in the extracted information. Recall is the ratio of the extracted correct information to the information in all samples.

3.3.1. mAP (mean Average Precision)

Because more than one person is detected in the pedestrian re-identification task, we use mAP as the final evaluation index, which reflects the property of the correct search results arranged at the top of the tested list. The area under the PR curve (Precision-Recall) is defined as AP (Average Precision). It reflects the relationship between the horizontal coordinate R and the vertical coordinate P. Furthermore, the larger the area under the PR curve, the better the model performance. The Precision rate and Recall rate are calculated as eq. (1) and eq. (2):

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{1}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{2}
\]

Among them, TP (True Positives) means the number of detection frames with IoU > 0.5 (Intersection over Union, IoU); FP (False Positives) is the number of detection frames that indicate IoU ≤ 0.5, which can also be said to be the number of redundant detection frames that detect the same real label; FN (False Negatives) indicates the number of detection frames that have not detected real labels.

3.3.2. CMC (Cumulative Matching Characteristics)

Another metric is CMC (Rank-k matching accuracy). Rank-n indicates that in the descending order of similarity of recognition results, the first n results contain the target.

The recognition rate is the ratio of the number of Rank-n to the number of query samples; it is the same as the commonly used top1 err or top5 err evaluation index, but here Rank-1 recognition rate represents the correct rate rather than the error rate. The relationship between the two is eq. (3).
3.3. mINP (mean inverse negative penalty)

In passenger retrieval tasks, the ranking of the correct target in the search list should be considered, especially for multi-view data. The ranking of hard-samples further determines the measurement of model evaluation. It may happen that the AP obtained by Rank List 1 under the same CMC is better than Rank List 2, but finding all the right matches takes much more effort. To solve this problem, Ye [61] designed a negative penalty (NP), in which the hardest sample with the correct match is considered.

$$N_P = \frac{R_{hard} - G_i}{R_{hard}}$$  \hspace{1cm} (4)

where $R_{hard}$ in eq. (4) represents the sort location of the hardest sample, and $G_i$ indicates the number of pairs matched correctly. The inverse of NP is defined as INP, the calculation of mINP is given by eq. (5).

$$mINP = \frac{1}{n} \sum_i (1 - N_P) = \frac{1}{n} \sum_i \frac{G_i}{R_{hard}}$$  \hspace{1cm} (5)

3.3.4. RS (Robustness Score)

To further evaluate the robustness for retrieval objects in long time span, Huang [62] defined a Robustness Score (RS) which is defined as eq. (6):

$$RS(P) = \frac{\sqrt{Score_L(P) \times Score_S(P)}}{|Score_L(P) - Score_S(P)| + 1}$$  \hspace{1cm} (6)

where $Score_L$ and $Score_S$ represent the long-term and short-term re-ID accuracy respectively; $P$ represents different evaluation indexes. For a given ReID model, the closer $Score_L$ and $Score_S$ are to each other, the greater the value of the RS indicator.

4. Pedestrian reidentifcation based on deep learning

Deep learning based Pedestrian ReID including feature learning and distance metric. For the feature learning approach, the model treats ReID as a classification or validation task, rather than directly determining the similarity between images.

Feature representation learning mainly involves classification loss and verification loss. Specifically, ID is often used as the basis for classification loss. ID Embedding network (IDE) [47][63] is a widely used baseline model. Through using ID of a pedestrian as a training label to train the model, it only needs to input one picture at a time. While the input of verification loss is a pair of pictures of pedestrians, the network learns whether they are the same person, which is equivalent to a binary classification problem. The verification model aims to learn the similarity of two pictures. Generally it uses end-to-end models such as Siamese network [64], Fig. 6 summarizes various ReID algorithms and potential future directions. In this part, we will introduce cutting-edge ReID methods mainly from the perspective of global and local features-based, video sequence-based, GAN-based and attention mechanism-based.
4.1. Global and local features-based method

4.1.1. Global features

Global feature refers to the extraction of the overall information of each picture, it does not have any spatial information. While Local feature refers to the extraction of a certain area in pedestrian image, and finally multiple local information is merged as the final feature.

In the early years of incorporating advanced deep learning methods into the area of PAR and pedestrian ReID, global feature was the well known choice. Global features can usually be extracted end-to-end, and the advantage lies in simple calculations. But the limitations include that noise regions may cause great interference to global features, and the misalignment of poses will also lead to poor global feature matching.

4.1.2. Horizontal slices

The partial feature extraction method based on deep learning includes image block [65,66,67] and the posture-based local feature [36][68]. There are many blocking methods of many images, such as uniform block [65] [67] and context sensing block [69], where uniform block can be divided into grid block [65] and horizontal slice [66]. This part will introduce several levels of horizontal slice.

In Gate Siamese [70], each image is characterized by a CNN, and local features are input to the LSTM network in order, automatically express the image final feature. Deep-Person [71] regarded the pedestrian as a sequence of body parts, and apply LSTM to take into account the contextual information between body parts. AlignedRelD [72] introduced Dynamically Matching Local Information (DMLI), using shortest path to find the most effective dynamic connection (Dynamic Time Warping). PCB [66] horizontally cut the input image into 6 pieces, extracts a partial feature for each line and finally concatenate the local characteristics. Integration convolutional neural network (ICNN) [73] jointly learned global and local features. Fan et al. [74] proposed SCPNet which discriminative features are learned using the space-channel relationship, each channel is assigned to focus on a specific space of the body.

4.1.3. Posture information

The motivation of local features based on gesture extraction is to make full use of pedestrian key points [36][68]. First we migrate the trained gesture estimation model to the pedestrian dataset, so obtain the positioning of the human body posture skeletal point, then extract local characteristics based on key points. Fig. 7 shows an effective approach under pose changes, i.e., fusing global and local features on extracting key points of the human body.

Spindle Net [36] firstly used human body structure information to facilitate feature learning, which employed feature decomposition and feature fusion based on body region information. Zheng et al. [33] introduced the PIE to describe pedestrian, the Pose-Box structure was used to align pedestrians, which is generated through pose estimation followed by affine transformations. Pose-driven Deep Convolutional [34] proposed an end-to-end architecture that use partial human cues to mitigate pose changes. Wei et al. [35] explicitly leveraged the local and global information of body to generate more robust representation. PABP [75] learned part-aligned representation, the body feature maps are connected to a bilinear ensemble layer and then fused into a single image descriptor.

4.1.4. Semantic segmentation

Semantic segmentation of the image is also applied to the weight recognition, the representative methods are as follows. Generating realistic images of persons is a task full of challenge as the complex interplay between the different image factors, such as the foreground, background and pose information. Ma et al. [88] aimed at generating such images based on a novel, two-stage reconstruction pipeline, they give a new idea for image prospect extraction. SPRReID [89] used additional semantic parsing branches to generate probability maps associated with different regions. Besides, the method of semantic segmentation can also be referred to DeepLabV2 [90], Mask R-CNN [91], etc.

4.1.5. Grid characteristics

IDLA [92] simultaneously learned a similarity metric for person ReID. The input to the network is pairs of images, higher layers compute relationships by calculating the difference of the eigenvalues of a certain size grid. PersonNet [93] employed a layer of computing neighborhood range differences across two input images to capture local relationship between patches. For the problem of people images of different sizes, DSR [94] used FCN to make the pixel-level features consistent.

In general, the grid feature is a relatively fine-grained physical area feature. Some early works extended the grid feature to a part feature to calculate the feature map difference of two images. The above typical researches used the grid feature to solve ReID Work, but due to the large amount of calculation and the lack of obvious performance, the grid feature is not very commonly used.

4.2. Video sequence-based method

Video re-identification (Video ReID) is also called as sequence re-identification, which refers to the use of a continuous sequence of
pedestrian pictures for pedestrian re-identification tasks. Generally, the characteristics of Video ReID include rich pedestrian posture changes, common occlusion phenomena, difficulty in fusion frame information, and inconsistent quality between frames, etc.

The earlier method of video ReID is to extract per-frame features, and then directly obtain the final feature through average pooling or maximum pooling. This is relatively simple, and the performance depends on the single frame. The difficulty of sequence re-identification lies in how to judge the quality of each frame and perform feature fusion of multiple frame features, how to generate the motion features for the sequence, and improve the computational efficiency.

Accumulative Motion Context (AMOC) [95] contained spatial motion sub-network and employed RNN to fuse the characteristic information of all frames. Aiming at how to integrate features, Deep Feature Guided Pooling (DFGP) [96] used PCA-based convolutional network (PCN) to extract features of each frame and give a new idea to take full advantage of the space-time information. Region-based Quality Estimation Network (RQEN) [97] mainly focused on the case of occlusion. Besides, Saha et al. [98] used residual network (ResNet) along with LSTM for feature extraction. Zhang et al. [99] introduced the concept of “mean-body” and defined an intra-video loss. Yu et al. [100] exploited unlabeled tracklets, Li et al. [101] proposed a new spatio-temporal attention model that automatically discovers adverse set of distinctive body parts. TCA-Net [102] considered the temporal alignment problem.

Research on video ReID across time and space has received increasing attention in recent studies. Table 7 compares several cross-temporal video re-recognition algorithm indicators published at the top conference.

### Table 7. Performance comparison of several cross-temporal video re-recognition algorithms on MARS dataset.

| Method         | Source | MARS Rank1 | MARS Rank5 | MARS Rank20 | mAP  |
|----------------|--------|------------|------------|--------------|------|
| CTL [76]       | CVPR 21| 91.4       | 96.8       | 86.7         | 86.0 |
| BiCnet-TKS [77]| CVPR 21| 90.2       | -          | -            | 86.0 |
| GRL [78]       | CVPR 21| 91.0       | 96.7       | 98.4         | 84.8 |
| ST-GEN [79]    | CVPR 20| 89.95      | 96.41      | 98.28        | 83.70|
| MGH [80]       | CVPR 20| 90.0       | 96.7       | 98.5         | 85.8 |
| MG-RAFA [81]   | CVPR 20| 88.8       | 97.0       | 98.5         | 85.9 |
| AP3D [82]      | ECCV 20| 90.7       | -          | -            | 85.6 |
| APA [83]       | ECCV 20| 90.2       | 96.6       | -            | 82.9 |
| TCLNet [84]    | ECCV 20| 89.8       | -          | -            | 85.1 |
| STRF [85]      | ICCV 21| 90.30      | -          | -            | 86.10|
| STMN [86]      | ICCV 21| 90.5       | -          | -            | 84.5 |
| STEP-Emb [87]  | ICCV 21| 90.8       | 97.1       | 98.8         | 87.0 |

### Table 8. The realization methods and targets of a typical GAN network.

| Method  | Basis     | Tricks                   | Target                   |
|---------|-----------|--------------------------|--------------------------|
| DCGAN  | GAN       | Label smooth             | Extended data            |
| CamStyle | CycleGAN | Label smooth             | Camera style deviation   |
| PTGAN  | CycleGAN + PSPNet [103] | Background segmentation | Domain dissimilarity     |
| SPGAN  | CycleGAN + Siamnet | Constrain mapping       | Domain dissimilarity     |
| PNGAN  | GAN + openpose [104] | Extract body pose       | pose-normalization       |

#### 4.3. GAN based method

Generally, the performance of training models on different datasets has certain differences. The changes in experimental effect caused by changes in the training datasets are likely to be caused by overfitting of the model. To solve such problems, GAN is widely used in data augmentation and expansion. GAN [105] could randomly generate sample images, CGAN [106] added conditional constraints to generate images on the basis of GAN, Pix2pix [107] and CycleGAN [108] could convert paired images and arbitrary images in the A and B domain respectively. Zheng et al. [109] applied GAN to create unlabeled samples to expand data, and used the label smoothing regularization for outliers (LSRO) to smooth ID tags. StyleGAN [110] and CamStyle [111] could achieve style transfer between any two cameras.

PTGAN [112], SPGAN [113] solved the obvious deviation between the data collected in different scenarios, improved the network trained on the A dataset in the B scene by generating the image of the former into the style of the latter. PNGAN [114] generated the image of the desired posture according to the input image and the target posture, the features extracted from multiple fixed poses of the same pedestrian make the model more robust. Several GAN-based methods are compared and summarized in Table 8. Fig. 8 shows the images generated by several GAN networks, in which CamStyle transfers between two cameras, PNGAN generates a variety of poses, and PTGAN realizes style conversion in different domains.
4.4. Attention mechanism-based method

Attention mechanism [115] is similar to the human method of selective signal processing by focusing attention on specific areas while ignoring other information. It can be divided into time attention mechanism and spatial attention mechanism. The former mainly focuses on which sequence is more important, while the latter mainly focuses on specific part. Song [116] used a mask to divide the complete image into background and pedestrian body regions, and used the triple loss supervised network to learn body region features and ignore background features. Li et al. [101] solved the problems of pedestrian occlusion and misalignment in video sequences by new proposed spatiotemporal attention model. HA-CNN [117] jointly learned soft pixel as well as hard regional attention. Xu et al. [118] introduced Attention-Aware Compositional Network (AACN) for ReID, which consists of PPA and AFC, which represent pose-guided and attention-aware feature, respectively. There is also a typical HydraPlus-Net [39], that multi-directionally provide the multi-level maps with attention to different layers.

5. Challenges and frontier research

With increasing public safety requirements, pedestrian re-identification often faces the task of cross-camera retrieval. The primary difficulty in solving this problem is the appearance change caused by the difference in viewing angle or cloth-changing in different scenes. At the same time, some fine-grained attribute information needs to be considered to distinguish different people. In summary, pedestrian re-identification is challenging in learning a discriminative and robust visual representation for changes in perspective. Specifically, pedestrian re-identification may face problems and difficulties such as limited training data, algorithm performance, including effectiveness and efficiency, domain gap, unconstrained environment, etc. There are also some frontier researches to explore new structure, loss function design, and effective training tricks, etc.

5.1. Network structure design

5.1.1. ReID specific architecture

Existing ReID methods usually make some architectural modifications on the basis of image classification, mostly follow the CaffeNet [119], VGG-16 [120] and ResNet-50 backbone, for example, SVDNet [121] used the Eigenlayer as the second last FC layer, PCB [66] set the last convolutional stripe to 1 to increase feature layer pixels. Researchers have also proposed some new networks designed specifically for ReID. Noting that the metric and the classification loss are inconsistent in an embedding space, BNNeck [122] added a BN layer after global pooling to separate them into two different feature spaces. Omni-Scale Network (OSNet) [123] designed a residual block consists of feature stream that detects features of certain size. Auto-ReID [124] involved a new reID search space and a new retrieval-based model, it combined a typical classification search space and a novel part-aware module. Faster-ReID [125] gave an All-in-One framework which learns multiple codes of different lengths in a single framework with a code pyramid structure and self-distillation learning. Besides, there are some other typical networks including FPNN [65], MLFN [126], BraidsNet [127] and Siamese network [128], AGW [129] provided a baseline and achieves competitive performance on cross-modality ReID tasks.

5.1.2. Loss function design

To search for pedestrians with the same probe ID in the gallery, pedestrian re-identification often needs to achieve similarity-based clustering effects. Therefore, unlike the frequently used loss functions such as cross entropy, the commonly used metric learning methods and loss functions in pedestrian re-identification research are ID loss [16][121], verification loss [92][130][131], triplet Loss [132, 133, 134, 135] as shown in Fig. 9, and contrastive loss [136], quadruplet loss [137], fusion of part loss and global loss [36], multiple loss combinations [138], circle loss [139], etc. In unsupervised learning, other non-parametric classification losses such as InfoNCE [140], ClusterNCE loss [142], OIM loss [143].

Classification loss: It is also well known as ID loss. The ID number of pedestrians in the train set is consistent with the network categories.
The feature layer is connected with a FC layer for classification, and the cross entropy loss is calculated through the softmax activation function.

\[ L_{id} = \frac{-1}{n} \sum_{i=1}^{n} \log(p(y|x_i)) \]  

(7)

As shown in eq. (7), where the label of the input image \( x_i \) is represented as \( y_i \), \( p \) is the probability of network prediction, \( n \) represents the count of training samples in per batch.

**Verification Loss:** It fuses two feature information to calculate a contrastive loss [70], and each time a pair of pictures are input into the same Siamese network to extract features. The contrastive loss optimizes the comparison of paired sample distances, it can be formulated by eq. (8)

\[ L_{ciun} = y d^2_{i_k} + (1-y)(a - d^2_{i_k}) \]  

(8)

Where \( y = 1 \) when \( i_k \) and \( i_l \) are the same, and \( y = 0 \) on the contrary. \( a \) is a margin parameter. \( d_{i_k,i_l} \) and \( d_{i_l,i_k} \) indicates the Euclidean distance between pairs of input features of \( i_k \) and \( i_l \), which can be expressed by eq. (9)

\[ d_{i_k,i_l} = \| f(i_k) - f(i_l) \|_2 \]  

(9)

**Triplet Loss:** Arising in the context of nearest neighbor classification [144], \( x \) is an image that is embedded into Euclidean space and constrained to a \( d \)-dimensional hypersphere \( \| x \|_2 \). For a specific person, to ensure that the anchor \( x_i \) is closer to the positive sample \( x_{i_p}^k \) than the negative \( x_{i_n}^k \), define the loss function as eq. (10).

\[ L = \sum_{i} \left[ \| f(x_i^k) - f(x_{i_p}^k) \|_2^2 - \| f(x_i^k) - f(x_{i_n}^k) \|_2^2 + \alpha \right] \]  

(10)

**Unified Loss:** Maximizing intra-class similarity \( s_p \), and minimizing inter-class similarity \( s_n \) are main purposes of deep feature learning. In the feature space, given a sample \( x \), assume that the similarity scores of within-class and between-class are \( K \) and \( L \). \( \gamma \) is a scale factor and \( m \) is a margin for better similarity separation. We denote these similarity scores as \( \{ s_p^i \}(i = 1, 2, \ldots, K) \) and \( \{ s_n^j \}(j = 1, 2, \ldots, L) \), respectively. \( L_{uni} \) iterates through every similarity pair to reduce \( s_p^i - s_n^j \). It degenerates to triplet loss or classification loss through slight modifications. To minimize each \( s_p \) as well as to maximize each \( s_n \) as well as to maximize \( s_p \), unified loss function is defined by:

\[ L_{uni} = \log \left[ 1 + \sum_{i=1}^{K} \sum_{j=1}^{L} \exp(y(s_p^i - s_n^j + m)) \right] \]  

(11)

**Circle Loss:** It allows each similarity score to be learned at their own speed, thus increasing the increasing of optimization, depending on its current optimization status. \( a_p^i \) and \( a_n^j \) are non-negative weighting factors. Firstly neglect the margin item \( m \) in eq. (11) and transfer the unified loss function into the proposed circle loss and get eq. (12).

\[ L_{circular} = \log \left[ 1 + \sum_{i=1}^{K} \sum_{j=1}^{L} \exp(y(a_p^i - a_n^j)) \right] \]  

(12)

**Softened Cross-entropy Loss:** Lin et al. [145] proposed softened similarity learning and retrieved the error of the hard quantization loss. Unlike original one-hot labels, an image is encouraged to be associated with \( \lambda \) possible related classes, whose labels are treat as a distribution. The softened cross-entropy loss is formulated as eq. (13).

\[ L = -\sum_{j=1}^{N} \log(p(y_j|x,V)) \]  

(13)

Where the probability of the \( i_{th} \) class is defined as \( p(y|x,V) \), \( V \) is the lookup table that stores the feature of each class. Hyper-parameter \( \lambda \) balances the effect of the ground truth and the reliable classes.
shape and skeleton information regardless of illumination and color change. Contour sketch [153] of person image was proposed to take advantage of the shape of the human body for extracting features. 3D information caught by RGB-D Sensors [154] was considered to be a very fruitful research direction because of its main advantage of a soft biometric policy. Huang et al. [62] adopted Feature Sparse Representation and Soft Embedding Attention, integrated capsules to deal with the person re-ID task [155]. There are also Cloth-Clothing Change Aware Network (CCAN) [156] which performed identity recognition and clothing change detection simultaneously. The above methods also established some typical datasets, including COCAS [151], RGB-D [154], PRCC [153], Celeb [157], VC-Clothes [150], DG-Net [158], some samples of which are shown in Fig. 11.

5.3. Domain adaptation

After the application scenario changes, the performance of the algorithm trained with a specific dataset could vary greatly. It is well known that there are large domain gaps among different datasets [64][159]. For example, in the data collected by the same camera in summer, short-sleeved shirts will account for more pedestrian attributes, and backpack attributes will be more common in the data collected by outdoor cameras. For the problem of insufficient data and limitation of scene area, the data can be augmented by synthesizing images through style transfer methods such as GAN [105].

Training across datasets to achieve domain adaptation is a common method used in existing researches. In addition, it is possible to train a domain generalized model that contains multiple source datasets, so it can be transferred to new datasets without additional learning. Domain Invariant Mapping Network [160] was proposed to enable a ReID model to be deployed out-of-the-box for any new camera network domain. There are also PUL [161] methods that don’t label the data of the new scene, but assign pseudo-labels to the new data through the method of clustering, and finally fine-tune the existing model. HHL [162] and Tracklet Association [163] used prior knowledge and soft label to further mine information in the target domain. Some progresses have also been made on more general classification and semantic segmentation [164] issues. Kang [165] aligned the attention maps of different domain so that the information is better adapted from source to the target network, which inspired a new way and made contributions to unsupervised domain adaptation (UDA).

5.4. Occlusion conditions

In actual scenes, occlusion is often impossible to avoid, only a portion of the human body is available [67], resulting in a decrease in the accuracy of the model, which requires the model to have a high generalization ability. The general idea to solve this problem in person recognition can be divided into two types. One is to use the key point detection [166] to extract the non-occluded part of the characterization for similarity matching, and another is to build a large-scale occlusion datasets, among which representative ones are Occluded-ReID [167], Partial-ReID [67] and P-ETHZ [168].

Using the posture information to extract the key points is a better approach to locate the occluded part of the body. OGS-Net [169] map the human body to 3D space, used the point cloud to integrate the structure to learn the human body expression and obtain robust features. PGFA [166] exploited marks of pose to disentangle the useful information from the occlusion noise. HO-Net [170] learned high-order relation information assisted by posture for robust alignment.

For partial re-identification, Visibility-aware Part Model (VPM) [168] perceived the visibility of regions through self-supervision, which was capable to estimate the shared regions between two images and thus suppresses the spatial mis-alignment. Deep Spatial feature Reconstruction DSR [94] used half-length images to search for partially occluded full-body images, and generated spatial feature maps with fixed size to remain features consistent at pixel level. FPR [171] was proposed for Alignment-free Occluded Person ReID. Aligned-ReID [172][173] used Dynamically Matching Local Information (DMLI) dynamic connection to solve horizontal block misalignment.

The performances of typical methods under various occlusion datasets are shown in Table 9. Besides, multi-modal input can also improve model robustness in unrestricted case, which makes the model closer to practical use, such as searching by verbal description. Textual-visual matching [197] measured similarities between sentence descriptions and images.

5.5. Model efficiency

Lightweight ReID model is getting more and more attention in current research. For higher accuracy, most of the existing methods utilize a large deep network to learn high-dimensional features for computing similarities. But the query time would increase massively with the expand of the tested gallery size.

The most direct idea to reduce model complexity is mainly focused on network pruning for image retrieval acceleration. Compared with the
Table 9. Performance comparisons with the occluded methods on the reported datasets.

| Method            | Occluded-REID | Partial-REID | P-DukeMTMC |
|-------------------|---------------|--------------|------------|
| IDE [63]          | 52.6          | 46.4         | 51.7       |
| PCB [66]          | 59.3          | 53.2         | 66.3       |
| PGFA [166]        | 57.1          | 56.2         | 68.0       |
| PVPM [174]        | 66.8          | 59.5         | 75.3       |
| Occluded-REID [167]| 68.14         | -            | 78.52      |
| VPM [175]         | -             | 67.7         | -          |
| DSR [84]          | 72.80         | 62.83        | 73.67      |
| FPR [171]         | 78.30         | 68.00        | 81.00      |

Fig. 12. State-of-the-arts (SOTA) of ReID on Market-1501 datasets.

conventional global center-based methods, which only keep the filters far away from the geometric center, Progressive Local Filter Pruning (PLFP) [198] pruned filters according to the local relationships to the neighbors to preserve the representation ability of the model. Based on a new search space and a new retrieval algorithm, Auto-ReID [124] enabled the automated approach to find an efficient and effective CNN architecture.

To search person images quickly and accurately. The main idea of recent fast ReID methods is hashing algorithm which learns a binary code instead of real value. Several fast ReID methods [199, 200, 201, 162, 202] have been proposed to increase algorithm speed while maintaining competitive accuracy, a typical example is CtF [125] search strategy. Another research direction is to design a lightweight network by modifying the model. [123][203].

5.6. Limited training data

Existing pedestrian datasets are generally obtained for common scenarios. In most cases, in order to generate more effective models, training specific networks require specific data sets. However, collecting and labeling is a very time-consuming and laborious work. In order to expand the limited dataset, the most popular approach in the research is GAN [105] and its derivative models including CamStyle [111], PTGAN [112], SPGAN [113], PNGAN [114] etc. Besides, APR [16] added annotated appearance attributes for the Market-1501 dataset and the Duke-MTMC dataset, making the dataset originally used for pedestrian ReID can be applied to attribute recognition. DGNet [158] used clothes-changing pedestrian images to enhance data. SOMANet [204], PersonX [205] and VC-Clothes [150] used 3D game engine to generate human data in different views. Huang [157] crawled the street snap-shots of celebrities from the Internet and built Celeb-ReID.

5.7. Resolution changes

Due to changes in camera angles and inconsistent target distances, the resolution of pedestrian pictures varies greatly. SING [206] simultaneously optimized image super-resolution and person ReID matching, which use different methods to extract features of different resolutions. DSPDL [207] designed a Discriminative Semi-coupled Projective Dictionary Learning framework for matching pedestrian images of great resolution divergences. SDALF [208] (Symmetry-Driven Accumulation of Local Features) consisted in the extraction of features that model three complementary aspects of the human appearance, in this way, robustness to viewpoint and illumination variations is achieved. A new work in CVPR2020 named as INTACT [209] (Inter-Task Association Critic) gave an effective solution to use an image super-resolution model to improve the low resolution of query images, which can reduce unaligned feature distributions and promote identity matching performance.

6. State-of-the-arts

6.1. Analysis of ReID on SOTA

This part mainly reviews the ReID works published in top conferences in computer vision and artificial intelligence, and makes a comparative analysis with some other representative works. As Table 10 shown, we summarize the outcomes mainly on Market-1501 and DukeMTMC datasets. Correspondingly, the visualized scatter plot is depicted in Fig. 12.
Table 10. Comparison with state-of-the-arts for ReID on Market-1501 and DukeMTMC.

| Method              | Source     | Market-1501 mAP | Market-1501 Rank1 | DukeMTMC mAP | DukeMTMC Rank1 |
|---------------------|------------|-----------------|------------------|--------------|----------------|
| Supervised Learning |            |                 |                  |              |                |
| Deep-Person [71]    | Pattern Recognition | 79.58 | 92.31 | 64.80 | 80.90 |
| AlignedReID + + [173] | Pattern Recognition | 77.6 | 91.0 | 68.0 | 80.7 |
| FMN [176]           | Pattern Recognition Letters | 67.12 | 85.99 | 56.88 | 74.51 |
| MGN [177]           | ACM Multimedia | 86.9 | 95.74 | 78.4 | 88.7 |
| PAN [178]           | TSVT       | 63.5 | 82.81 | 51.51 | 71.59 |
| GF-reid [179]       | ArXiv 18 | 81.2 | 92.2 | 72.8 | 85.2 |
| OG-Net [169]        | ArXiv 20 | 59.97 | 80.94 | 50.81 | 71.77 |
| ACRN [21]           | CVPR 17 | 62.60 | 83.61 | 51.96 | 72.58 |
| SVDNet [121]        | CVPR 17 | 62.1 | 82.3 | 56.8 | 76.7 |
| AACN [118]          | CVPR 18 | 66.87 | 85.90 | - | 41.37 |
| SReID [80]          | CVPR 18 | 79.67 | 91.45 | 68.78 | 83.3 |
| HA-CNN [117]        | CVPR 18 | 75.7 | 91.2 | 63.8 | 80.5 |
| HOReID [170]        | CVPR 20 | 84.9 | 94.2 | 75.6 | 86.9 |
| CamStyle [111]      | CVPR 18 | 71.55 | 89.49 | 57.61 | 78.32 |
| PCB [66]            | ECCV 18 | 77.3 | 92.4 | 65.3 | 81.9 |
| PISNet [180]        | ECCV 20 | 87.1 | 95.6 | 78.7 | 88.8 |
| Unsupervised Learning |          |                 |                  |              |                |
| UnityStyle [181]    | CVPR 20 | 95.8 | 98.5 | 93.6 | 95.1 |
| M² [182]            | CVPR 20 | 82.6 | 5.4 | 68.5 | 84.7 |
| Unsupervised re-ID [183] | CVPR 20 | 37.8 | 71.7 | 28.6 | 52.5 |
| SADA [184]          | CVPR 20 | 59.8 | 83 | 55.8 | 74.5 |
| Generalizing ReID [185] | ECCV 20 | 71.5 | 88.1 | 65.2 | 79.5 |
| MEB-Net [186]       | ECCV 20 | 76.0 | 89.9 | 66.1 | 79.6 |
| CrowdReID [187]     | ECCV 20 | 84.7 | 95.3 | 74.4 | 88.3 |
| Light-ReID [125]    | ECCV 20 | 84.9 | 93.7 | 74.8 | 87.6 |
| UNRN [188]          | AAAI 21 | 78.1 | 91.9 | 69.1 | 82.0 |
| Transfer Learning   |            |                 |                  |              |                |
| SPGAN [113]         | CVPR 18 | 26.9 | 58.1 | 26.4 | 46.9 |
| VPM [175]           | CVPR 19 | 80.8 | 93.0 | 72.6 | 83.6 |
| HCT [189]           | CVPR 20 | 56.4 | 80.0 | 50.7 | 69.6 |
| SCRN [190]          | CVPR 20 | 88.5 | 95.7 | 79 | 91 |
| AD-Cluster [191]    | CVPR 20 | 68.3 | 86.7 | 54.1 | 72.6 |
| HHL [162]           | ECCV 18 | 31.4 | 62.2 | 27.2 | 46.9 |
| BUC [192]           | AAAI 18 | 38.3 | 66.2 | 27.5 | 47.4 |
| ARN [193]           | CVPR 18 | 39.4 | 70.3 | 33.4 | 60.2 |
| SSG [194]           | ICCV 19 | 58.3 | 80.0 | 53.4 | 73.0 |
| MMCL [195]          | CVPR 20 | 60.4 | 84.4 | 51.4 | 72.4 |
| NRMT [196]          | ECCV 20 | 71.7 | 87.8 | 62.2 | 77.8 |

Analyzing the improvement effect of the above new algorithms on the model performance, it can be seen that some tricks have brought objective efficiency, as discussed below. Applying of GAN models (e.g., DI-ReID [210] and UnityStyle [181]) could obtain more style-robust depth features for querying. Domain adaptive person ReID, including SADA [184], AD-Cluster [191], MEB-Net [186] and Generalizing ReID [185] gradually become a research hotspot. Besides, unsupervised ReID [183] adopt iterative training. PRI [211] solved low-resolution problem. Ahmed et al. [212] used transfer learning that aims to transfer the knowledge using only source models and limited labeled data. HCT [189] conducted training with hard-batch triplet loss. There are also novel methods that focus on solving some challenges, representatives are HORereID [170], PISNet [180] and CrowdReIDGASM [187], which studied ReID under occlusion, pedestrian interference and crowded conditions respectively.

For cross-domain re-identification, as Table 10 displayed, we evaluated some cross-domain ReID methods published on top conferences on two public person ReID datasets: Market-1501 and DukeMTMC. Quantitatively, we compared these state-of-the-art domain adaptation person ReID approaches. The comparison shows the influences on Map and Rank-1,5,10(%), with Market-1501 and DukeMTMC-ReID as source and target datasets, respectively.

Most work published on 2020 ECCV focused on unsupervised ReID, especially unsupervised domain adaptation (UDA), as the methods listed in the table. Typically, GDS [213] and JVTC [214] designed new training strategies. DG-Net + [215] segregated the id-unrelated noises to get more effective adaptation from the adaptation process. AD-Cluster [191] introduced adaptive sample augmentation to generate more diverse samples.

6.2. Potential research direction of ReID

Domain Adaptation: According to Table 10, it can be seen that domain adaptive re-recognition has been extensively studied in recent years, but it is still full of challenges to achieve rank-1 above 80% when the model trained under a specific dataset is transferred to others or applications in real scenarios or even unseen domains. Therefore, ReID in rich scenarios is a promising research direction. In particular, most datasets are collected in scenes with a wide field of view, so “partial” pedestrian re-recognition technology is very important.

Visible-Infrared ReID: Pedestrian re-identification at night or under poor light conditions is also a problem worthy of study, which is an important task in night-time surveillance applications. The main idea to solve this problem is to use infrared cameras to collect pictures. Hi-CMD [216], cm-SSFT [217], CAMalign [218] and DDAG [219] provided new ideas for cross-modality person re-identification respectively.

Unsupervised and transfer learning: Since the collected data is limited in most cases, the workload of manual labeling is huge and
costly. Semi-supervised [220], unsupervised learning [221] methods and transfer learning [222] have achieved widespread attention in recent studies, which also reflected by the increasing amount of work published in top venues. There has been some research works in this direction [183, 223, 224], but there is still a lot of rooms for improvement and full of challenges.

**High-quality datasets:** Most of the current pedestrian re-identification approaches are evaluated based on well-defined datasets. In practical scenarios, the ReID data might be collected from a variety of heterogeneous modalities. Due to domain differences, only the establishment of a large enough high-quality standard dataset can better train the robustness of the algorithm. For some clothing-change occasions, the author has created new datasets [150, 151, 153, 12], human interaction or active learning [225, 226] also provides another possible way to reduce the dependence on manual annotation.

6.3. Analysis of Attribute-Assisted ReID on SOTA

Pedestrian can be described by a combination of various different kinds of attributes, which can be a decisive clue to distinguish people with very similar global appearance. Through the Attribute-Assisted ReID algorithms mentioned in chapter 2.1, it can be drawn that person attribute recognition plays an auxiliary role in re-identification.

PAR and ReID can transfer information by attributes sharing. There are two ways to integrate attribute and ID information: (a) Independent Supervision [20]: Independently train a deep CNN for either attribute or identity label then use the concatenated feature for ReID matching. (b) Joint Supervision [27]: Multi-task joint learning CNN subjecting the identity and attribute supervision to a shared feature representation in the end-to-end model training.

Table 11, 12 show the performance of several Attribute Assisted ReID algorithms on Market-1501/DukeMTMC dataset and VIePR dataset, respectively. Experimental results demonstrated the effective ReID performance gain obtained by attribute assisting [11][15]. At the same time, the integration of ID effectively introduces complementary information to some extent and helps to obtain the improvement of overall attribute identification [19][16].

Analyzing the attribute-assisted ReID method, we can draw the following conclusions:

1) **Attribute attention model** generate features by “seeing” and “comparing” people images to locate the most discriminative parts [14], which provide new clues for searching for target persons given appearance descriptions [17].

2) **Attribute consistency principle** can be exploited to achieve a clear advantage in unsupervised fusion of multiple supervisions for cross-domain ReID [27], as well as in addressing human misalignment caused by pose changes [11].

3) **Attribute semantic integration model** jointly learns the discriminative projection of the appearance attribute subspace can effectively assist ReID [23].

4) **Attribute dataset fine-tuning** improves CNN features, combined with metric learning on ReID dataset can further improve the discriminative ability. [12][21][22].

Table 11. Comparison of several state-of-art Attribute-Assisted ReID methods on Market-1501 and DukeMTMC-ReID datasets.

| Method         | Market-1501 | DukeMTMC |
|----------------|-------------|----------|
|                | Map | Rank1 | Map | Rank1 |
| TJ-AIDL [27]   | 26.5 | 58.2 | 23.0 | 44.3 |
| ASMR [26]      | 31.0 | 49.6 | -   | -    |
| ACRN [21]      | 62.60 | 83.61 | 51.96 | 72.58 |
| APR [16]       | 66.89 | 87.04 | 55.56 | -    |
| A^3 M [8]      | 68.97 | 86.54 | -   | -    |
| APDR [11]      | 80.1 | 93.1 | -   | -    |
| Tensor Fusion  | 90.16 | 98.38 | 84.74 | 95.42 |

Table 12. Comparison of the accuracy and re-identification performance of several state-of-art Attribute-Assisted ReID methods on VIePR datasets.

| Method         | VIePR |
|----------------|-------|
|                | Acc | R1 | R5 | R10 | R20 |
| c-LSVM [28]    | -   | 17.09 | 38.61 | 52.53 | 68.35 |
| SSDAL [33]     | 58.6 | 37.9 | 65.5 | 75.6 | 85.4 |
| OAR [19]       | 66.9 | 41.5 | 69.0 | 82.5 | -   |
| CAN [14]       | -   | 54.1 | 83.1 | 91.8 | 96.4 |
| FT-CNN [42]    | -   | 52.1 | 79.6 | 89.2 | 95.0 |

Fourthly, training datasets can be expanded using pose or style transfer methods, which can effectively solve uncontrollable environmental factors such as domain deviation and clothing change. In general, with the booming development of deep learning technology and the urgent pace of building an intelligent society, re-identification and attribute recognition are facing many challenges and have significant potential research value.

**Declarations**

**Author contribution statement**

All authors listed have significantly contributed to the development and the writing of this article.

**Declaration of interests statement**

The authors declare no conflict of interest.

**Data availability statement**

Data will be made available on request.

**Funding statement**

Jie Hu was supported by National Social Science Fund of China [17ZDA020], National Natural Science Foundation of China [51975360 & 52035007]. This work was supported by Cross Fund for medical and Engineering of Shanghai Jiao Tong University (YQ2021QN118).

**Additional information**

No additional information is available for this paper.

**References**

[1] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2528–2535.

[2] N. Martinel, C. Micheloni, G.L. Foresti, Saliency weighted features for person re-identification, in: European Conference on Computer Vision, Springer, 2014, pp. 191–208.

[3] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012) 653–668.
[4] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 1, IEEE, 2005, pp. 886–893.

[5] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, IEEE, 1999, pp. 1150–1157.

[6] C.-F. Chao, F. Li, T.S. Huang, L. Cao, J.R. Smith, Learning locally-adaptive decision functions for person verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3610–3617.

[7] D. Wu, H. Huang, J. Qi, G. Xue, Overview of deep learning based pedestrian attribute recognition and re-identification, 2022, available at arXiv:2002891.

[8] H. Jia, L. Guo, Z. Zhang, A metric learning model for fine-grained representation learning, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 2040–2048.

[9] R. Layne, T.M. Hospedales, S. Gong, Q. Mary, Person re-identification by attributes, in: BMVC, vol. 2, 2012, p. 8.

[10] C. Lu, H. Zhao, S. Mian, T. Shao, J. Shao, S. Yi, Y. Xu, X. Wang, SpindleNet: person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.

[33] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3960–3969.

[34] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, GLAD: global-local-attention descriptor for pedestrian retrieval, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 423–431.

[35] H. Zhao, M. Tian, S. Sun, J. Shao, Y. Yan, S. Yi, X. Wang, TongueNet: person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.

[36] W. Deng, P. Liu, C.C. Loy, Person re-identification by pedestrian attribute recognition at far distance, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 789–792.

[37] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3774–3782.

[38] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, HydraPlusNet: attentive deep features for pedestrian analysis, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[39] W. Li, R. Zhao, T. Xiao, X. Wang, DeepSIFT: 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.

[40] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 79–88.

[41] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, re-identification, and tracking, in: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, 2007, pp. 1–7.

[42] L. Zhang, C. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, MARS: a video benchmark for large-scale person re-identification, in: European Conference on Computer Vision, Springer, 2016, pp. 868–884.

[43] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by discriminative selection in video ranking, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 2501–2514.

[44] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5177–5186.

[45] G. Song, B. Leng, Y. Liu, C. Hetaq, S. Cai, Region-based quality estimation network for large-scale person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[46] L. Zhang, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1367–1376.

[47] C.C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, in: 2013 IEEE International Conference on Image Processing, IEEE, 2013, pp. 3567–3571.

[48] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: Asian Conference on Computer Vision, Springer, 2012, pp. 31–44.

[49] W. Li, X. Wang, Locally aligned feature transforms across views, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3594–3601.

[50] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2009) 1627–1645.

[51] A. Dos, A. Chakraborty, A.K. Roy-Chowdhury, Consistent re-identification in a camera network, in: European Conference on Computer Vision, Springer, 2014, pp. 330–345.

[52] L. Ma, H. Liu, L. Hu, C. Wang, S. Sun, Orientation driven bag of appearances for person re-identification, arXiv preprint, arXiv:1605.02464, 2016.

[53] A. Dehghan, S. Modiei Anvari, M. Shah, GMM-GPCR: globally optimal generalized maximum multi clique puzzle problem for multiple object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4091–4099.

[54] M. Goo, Z. Wu, A. Rates-Borras, O. Camps, R.J. Radke, et al., A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 523–536.

[55] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1532–1545.

[56] M. Goo, S. Karanam, W.I. Jolliffe, R.J. Radke, DuketMTCAR4D: a large-scale multi-camera person re-identification dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 10–19.

[57] R. Benenson, M. Omran, J. Hosang, B. Schiele, Ten years of pedestrian detection, what have we learned?, in: European Conference on Computer Vision, Springer, 2014, pp. 613–627.

[58] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: European Conference on Computer Vision, Springer, 2014, pp. 688–703.

[59] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3958–3967.

[60] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
D. Wu, H. Huang, Q. Zhao et al.

[92] E. Z. RoCal

Temporal parsing

ECCV

Feature

Architecture

in: Aich, P.

L. Li, Jin,

Rao,

S.

re-identification, Papandreou,

Computer

T.

International

B.

F.

W. Tang,

re-identification.

re-identification.

re-identification.

re-identification.

re-identification, Multidata

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification.

re-identification, Video-based person re-identification by re-ranking attentive temporal information in deep recurrent convolutional networks, 2018, pp. 1663–1667.

re-identification, 2017.

re-identification.

re-identification.

re-identification, Video-based person re-identification by deep feature-guided pooling, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

re-identification, Region-based quality estimation network for large-scale person re-identification, 2017.

re-identification.

re-identification, Video-based person re-identification by re-ranking attentive temporal information in deep recurrent convolutional networks, 2018, pp. 3696–3699.

re-identification, Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3219–3228.

re-identification, European Conference on Computer Vision, 2016.

re-identification, Person re-identification, Pattern Recognition, 98 (2020) 107036.

re-identification, on human-level performance on fine-grained classification, 2017.

re-identification, in camera networks, IEEE Access (2018) 1.

re-identification, for video person re-identification, 2021.

re-identification, Learning efficient spatial-temporal representation for video person re-identification, 2021.

re-identification, Watching you: global-guided reciprocal learning for video-based person re-identification, 2020.

re-identification, for video person re-identification, 2020.

re-identification, Appearance-Preserving 3D Convolution for Video-based Person Re-identification, 2020.

re-identification, Temporal coherence or temporal motion: which is more critical for video person re-identification?, in: Computer Vision-ECCV 2020, 2020.

re-identification, Temporal complementary learning for video person re-identification, 2020.

re-identification, Spatial-temporal representation factorization for video-based person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 152–162.

re-identification, Video-based person re-identification with spatial and temporal memory networks, 2021.

re-identification, Dense interaction learning for video-based person re-identification, 2021.

re-identification, Disentangled person image generation supplementary material.

re-identification, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, 2020.

re-identification, Appearance-Preserving 3D Convolution for Video-based Person Re-identification, 2020.

re-identification, Temporal coherence or temporal motion: which is more critical for video person re-identification?, in: Computer Vision-ECCV 2020, 2020.

re-identification, Temporal complementary learning for video person re-identification, 2020.

re-identification, Video-based person re-identification with spatial and temporal memory networks, 2021.

re-identification, Dense interaction learning for video-based person re-identification, 2021.

re-identification, Disentangled person image generation supplementary material.

re-identification, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, 2020.

re-identification, Appearance-Preserving 3D Convolution for Video-based Person Re-identification, 2020.

re-identification, Temporal coherence or temporal motion: which is more critical for video person re-identification?, in: Computer Vision-ECCV 2020, 2020.

re-identification, Temporal complementary learning for video person re-identification, 2020.

re-identification, Video-based person re-identification with spatial and temporal memory networks, 2021.

re-identification, Dense interaction learning for video-based person re-identification, 2021.

re-identification, Disentangled person image generation supplementary material.

re-identification, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, 2020.

re-identification, Appearance-Preserving 3D Convolution for Video-based Person Re-identification, 2020.

re-identification, Temporal coherence or temporal motion: which is more critical for video person re-identification?, in: Computer Vision-ECCV 2020, 2020.

re-identification, Temporal complementary learning for video person re-identification, 2020.

re-identification, Video-based person re-identification with spatial and temporal memory networks, 2021.

re-identification, Dense interaction learning for video-based person re-identification, 2021.

re-identification, Disentangled person image generation supplementary material.

re-identification, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, 2020.

re-identification, Appearance-Preserving 3D Convolution for Video-based Person Re-identification, 2020.

re-identification, Temporal coherence or temporal motion: which is more critical for video person re-identification?, in: Computer Vision-ECCV 2020, 2020.

re-identification, Temporal complementary learning for video person re-identification, 2020.

re-identification, Video-based person re-identification with spatial and temporal memory networks, 2021.

re-identification, Dense interaction learning for video-based person re-identification, 2021.

re-identification, Disentangled person image generation supplementary material.

re-identification, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, 2020.

re-identification, Appearance-Preserving 3D Convolution for Video-based Person Re-identification, 2020.

re-identification, Temporal coherence or temporal motion: which is more critical for video person re-identification?, in: Computer Vision-ECCV 2020, 2020.

re-identification, Temporal complementary learning for video person re-identification, 2020.

re-identification, Video-based person re-identification with spatial and temporal memory networks, 2021.
A. Wu, H. Huang, Q. Zhao et al.

Recognition, identification, and re-identification, Pattern Recognition, 2018.

Z. Li, X. Pan, H. Wu, Improved feature learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

F. Zhao, J. Liu, H. Jiang, Improved face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

H. Huang, Q. Wu, J. Xu, Y. Zhong, Celebrities-ReID: a benchmark for clothes variation in long-term person re-identification, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019.

Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

S. Liao, L. Shao, Interpretable and generalizable person re-identification with query-agnostic convolution and temporal lifting, 2019.

J. Song, Y. Yang, Y.Z. Song, T. Xiang, T.M. Hospedales, Personalized person re-identification by domain-invariant mapping network, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: clustering and fine-tuning, ACM Trans. Multimedia Comput. Commun. Appl. 14 (2018) 1–18.

Z. Zheng, L. Zheng, S. Li, Y. Yang, Generalizing a Person Retrieval Model Hetero- and Homogeneously, Springer, Cham, 2018.

H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification by deep learning triplet association, ACM Trans. Multimedia Comput. Commun. Appl. 14 (2018) 1–18.

Z. Zheng, Y. Yang, Unsupervised scene adaptation with memory regularization in vivo, arXiv preprint, arXiv:1912.11164, 2019.

G. Kang, L. Zheng, Y. Yan, Y. Yang, Deep Adversarial Attention Alignment for Unsupervised Domain Adaptation: the benefit of Target Expectation Maximization, Springer, Cham, 2018.

J. Miao, Y. Wu, P. Liu, Y. Ding, Y. Yang, Pose-guided feature alignment for occluded person re-identification, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

J. Zhou, Z. Chen, J. Lai, G. Wang, Occluded person re-identification, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2018, pp. 1–6.

A. Feng, A mobile vision system for robust multi-person tracking, 2008, pp. 1–8.

Z. Zheng, Y. Yang, Person re-identification in the 3D space, arXiv preprint, arXiv: 2006.04569, 2020.

G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, High-order information matters: learning relation and topology for occluded person re-identification, 2020.

L. He, Y. Wang, W. Liu, X. Liao, H. Zhao, Z. Sun, J. Feng, Foreground-aware pyramidal reconstruction for alignment-free occluded person re-identification, IEEE, 2019.

Z. Xuan, L. Hao, F. Xing, W. Xiang, S. Jian, AlignedReID: surpassing human-level performance in person re-identification, 2017.

H. Luo, W. Jiang, X. Zhang, X. Fan, C. Zhang, AlignedReID++: dynamically matching local information for person re-identification, Pattern Recognit. 94 (2019).

S. Gao, J. Wang, H. Lu, Z. Liu, Pose-guided visible part matching for occluded ReID, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, J. Sun, Perceive where to focus: learning visibility-aware part-level features for partial person re-identification, IEEE, 2020.

G. Ding, S. Khan, Z. Tang, F. Porikli, Feature mask network for person re-identification, Pattern Recognit. Lett. 137 (2020) 91–98.

G. Wang, Y. Yuan, X. Chen, J. Ji, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 274–282.

Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Trans. Circuits Syst. Video Technol. 29 (2018) 3037–3045.

J. Almazan, B. Gajej, N. Murray, D. Larlus, Re-ID done right towards good practices and person re-identification, arXiv preprint, arXiv:1801.05339, 2018.

S. Zhao, C. Gao, J. Zhang, H. Cheng, X. Sun, Do Not Disturb Me: Person Re-identification Under the Interference of Other Pedestrians, 2020.

C. Liu, X. Chang, Y.D. Shen, Unity style transfer for person re-identification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

J. Zhou, B. Su, Y. Wu, Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

Y. Lin, L. Xie, Y. Wu, C. Yan, Q. Tian, Unsupervised person re-identification via soft-matching similarity learning, IEEE Trans. Multimedia Comput. Commun. Appl. 2020, pp. 3390–3399.

F. An, W. Jiang, H. Luo, M. Fei, SphereReID: deep hyperplane manifold embedding for person re-identification, 2018.

Z. Zheng, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, Proc. AAAI Conf. Artif. Intell. 34 (2017).

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, IEEE, 2016, pp. 2818–2826.

H. He, Z. Zhang, H. Zhang, Z. Jie, Light(B) Bag of tricks for image classification with convolutional neural networks, 2018.

F. Wan, Y. Wu, Q. Xian, Y. Chen, Y. Fu, When person re-identification meets changing clothes, IEEE, 2020.

S. Yu, S. Li, D. Chen, R. Zhao, J. Yan, Y. Qiao, COCAS: a large-scale clothes changing person dataset for re-identification, 2020.

A. Wu, W.-S. Zheng, J.-H. Lai, Robust depth-based person re-identification, IEEE Trans. Image Process. 26 (2017) 2588–2603.

Q. God, Ancong Wu, Wei-Shi Zheng, Person re-identification by contour sketch under moderate clothing change, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

I.B. Barbosa, M. Cristani, A.D. Rue, L. Bazzani, V. Murino, Re-identification with RGB-D sensors, in: International Conference on Computer Vision, 2012.

Z. Peng, W. Qiang, J. Xu, Z. Jian, Long-term person re-identification using true motion from videos, in: IEEE Winter Conference on Applications of Computer Vision, 2016.

X. Jia, Z. Meng, K. Katipally, H. Wang, K.V. Zon, Clothing change aware person identification, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.
