Multimedia Tools and Applications (2022) 81:21939–21973
https://doi.org/10.1007/s11042-022-12510-1

Survey for person re-identification based on coarse-to-fine feature learning

Minjie Liu[1,2] · Jiaqi Zhao[1,2] · Yong Zhou[1,2] · Hancheng Zhu[1,2] · Rui Yao[1,2] · Ying Chen[1,2]

Received: 13 July 2021 / Revised: 25 October 2021 / Accepted: 25 January 2022 /
Published online: 17 March 2022

Abstract

Person re-identification (Re-ID), aiming to retrieve interested people through multiple non-overlapping cameras, has caused concerns in pattern recognition communities and computer vision in recent years. With the continuous promotion of deep learning, the research on person Re-ID is more and more extensive. In this paper, we conduct a comprehensive review of the advanced methods and divide them into three categories from coarse to fine: (1) global-based methods, which are based on whole images to obtain discriminative features; (2) part-based methods, which focus on image regions to extract detailed information; (3) multiple granularities-based methods, which combine advantages of the above two categories. For each category, we further classify it according to popular research tools. Then, we give the evaluation of some typical models on a set of benchmark datasets and compare them in detail. We also introduce some widely used training tricks. The methods mentioned in this paper were published in 2011-2021. By discussing their advantages and limitations, we provide a reference for future works.

✉ Jiaqi Zhao
jiaqizhao@cumt.edu.cn

Minjie Liu
liuminjie@cumt.edu.cn

Yong Zhou
yzhou@cumt.edu.cn

Hancheng Zhu
hanchengzhu@cumt.edu.cn

Rui Yao
ruiyao@cumt.edu.cn

Ying Chen
cheny@cumt.edu.cn

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China

[2] Mine Digitization Engineering Research Center of Ministry of Education of the People's Republic of China, Xuzhou, 221116, China

Keywords Person Re-ID · Deep learning · Video surveillance

1 Introduction

With the continuous improvement of monitoring facilities and people's increasing safety awareness, more and more public places, especially schools, shopping malls, airports, and other places with a dense flow of people, begin to pay attention to the popularization of the monitoring network. These monitoring networks all over the world produce massive pictures or videos every day, which requires a lot of manual analysis and processing [1, 6, 77].

In response to this problem, various research began to focus on the intelligent analysis of the monitoring data. As two research hotspots in this field, person re-identification (Re-ID) and face recognition have caused concerns in pattern recognition communities and computer vision in recent years. Face recognition, on the basis of human face information, is a kind of biometric technology for identity re-identification [41, 42, 67]. It is widely used in access control systems, forensic scenarios and object tracking, ect [38–40]. Person Re-ID has important applications in criminal tracking, intelligent security and intelligent search [50, 73], ect. Given a query person of interest, the task of person Re-ID is to distinguish whether the person's images exist in the image gallery captured by other cameras [2, 34], as shown in Fig. 1. With the continuous promotion of deep learning, the research on person Re-ID is more and more extensive, among which the most in-depth research is the closed-world person Re-ID. The closed-world person Re-ID is based on the assumption that the queried image must consist in the gallery set and the closed-world scenarios, such as correct annotations, people represented by bounding boxes, and images captured by single modality cameras [53]. A typical person Re-ID model based on closed-world condition is shown in Fig. 2. Generally, a complete person Re-ID system is composed of three tasks: a person detection task, a person tracking task, and a person retrieval task [115].

The traditional person Re-ID methods are mainly divided into two categories: metric learning methods and feature learning methods. The metric learning methods generally

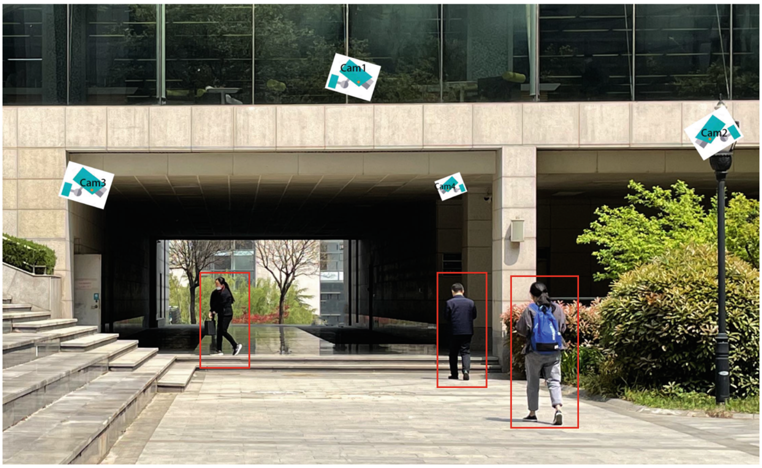


Fig. 1 Person images are captured by multi cameras

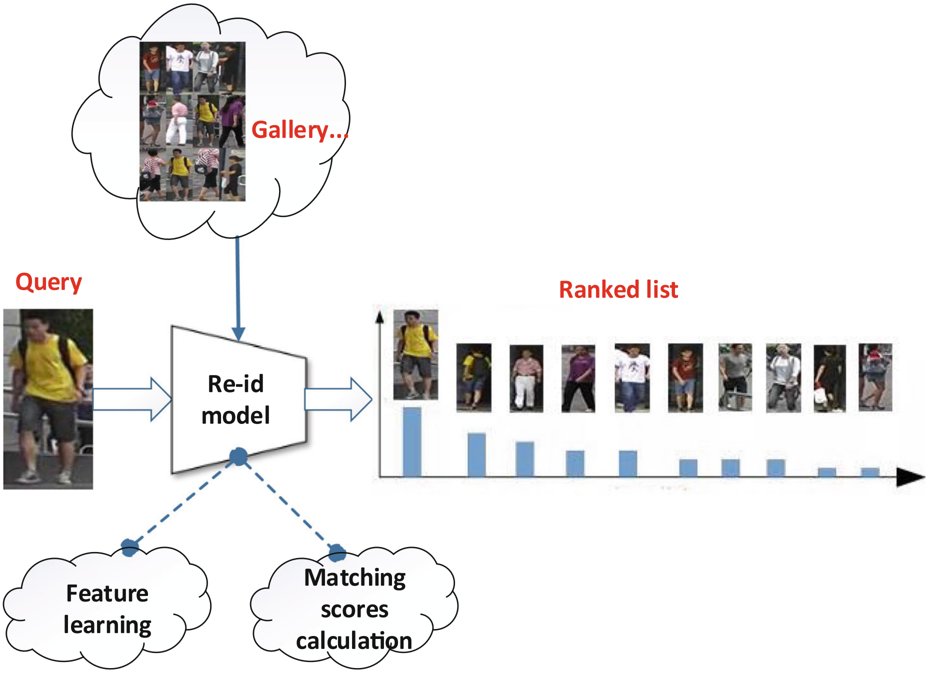


Fig. 2 Diagram of a person Re-ID model. Given a query image and a set of gallery images, the goal of a Re-ID model is to learn the features of images, calculate their similarity and generate a ranked list for discrimination

choose to learn an effective distance measure to impel the distance between features from different groups longer, and features from the same group closer [8, 47]. The methods based on image feature learning generally use manually designed features that are less sensitive to pose, illumination, and camera views [52, 111]. However, with the rapid development of deep learning, a growing number of methods combine the research of person Re-ID with deep learning and have achieved quite good results. Deep learning not only helps to extract high-level features but also innovates the research on metric learning. Methods combining deep learning are generally end-to-end, which realize the combination of metric learning and feature learning. In this paper, we review the recent person Re-ID methods based on deep learning. As a newly developing computer vision technology, person Re-ID faces several challenges in practical scenarios, as shown in Fig. 3. Firstly, images captured by surveillance cameras are usually in a complex environment and are generally of low pixels. Secondly, the image quality captured by different devices is uneven. These questions lead to some traditional biological features, such as facial and gait features, which are inadequate for the Re-ID task [109]. In this case, it seems more suitable for a person Re-ID task to extract people's appearance features from the color of their clothes or belongings. However, due to the change of illumination and the difference of perspective, the color itself often has great changes. Thirdly, the changes of pedestrian's poses are uncontrolled, which leads to the appearance features of even the same pedestrian are quite various in different time periods [35]. Fourthly, complex background environments and various occlusions also have negative effects on the extraction of appearance features [152]. Sometimes pedestrians may wear similar clothes, which often leads to the appearance of different pedestrians are

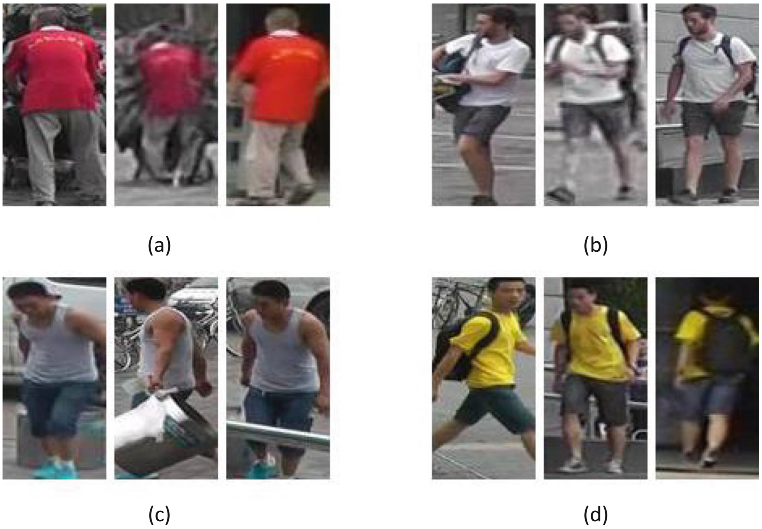


Fig. 3 Some challenges of person Re-ID. **a** Illumination variation; **b** Image quality difference; **c** Occlusions; **d** Pose variation. Images are randomly collected from Market1501

highly similar and increase the difficulty of Re-ID. In addition, the complex network structure brings huge computational cost, which makes some research on pedestrian recognition difficult in practical use [13, 107, 111, 117, 138].

In this work, we classify the research methods from another perspective. We conduct a comprehensive review of the mainstream research methods, and divide them into three categories from coarse to fine, according to the scale of the features studied: (1) global-based methods; (2) part-based methods; (3) multiple granularity-based methods. Here, 'coarse-to-fine' refers to the scale of feature learning, namely the mentioned three classifications. Similarly, the concept of 'coarse-to-fine' is also mentioned in paper [18]. In paper [18], 'coarse-to-fine' refers to the combination of global feature learning and local feature learning, which is classified as multiple granularity-based methods for person Re-ID in our paper. In addition, we further divide each classification according to the popular research tools. For global-based methods, we further divide it into three branches: attention-driven methods, image generation methods and attribute mining methods.

For part-based methods, we further divide it into three branches: image partition methods, feature description methods, and attention-driven methods. For multiple granularity-based methods, we further divide it into two branches: image partition methods and attention-driven methods. Based on the above three aspects, we summarize these methods in detail respectively, as shown in Fig. 4. The global-based methods are based on the overall situation and have relatively simple network structure, but ignore some important details. The part-based methods have poor integrity, but they can learn fine-grained information ignored by the global-based methods. The multiple granularity-based methods have the advantages of both the part-based methods and the global-based methods, but these methods are often accompanied by complex network structure and considerable computing consumption. Attention-driven methods can force the model to focus on discriminative salient features, while ignoring some sub-salient features. These neglected sub-salient features also have an important impact on Re-ID. Image generation methods can provide rich

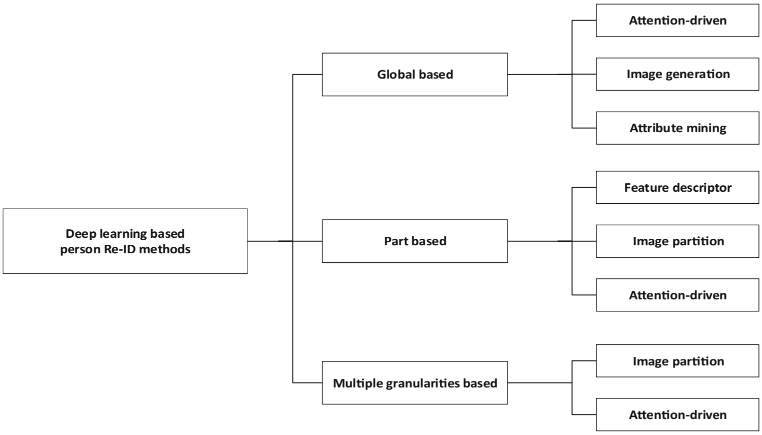


Fig. 4 Learning categories for Re-ID models. The papers mentioned for person Re-ID were published in 2011-2021

and diverse training samples, but there is still a certain gap between the generated image and the real person image, and they will lead to huge calculation consumption. In practice, it is not desirable to collect a large number of images of each pedestrian, and attribute learning can effectively overcome the problem of scarcity of category samples. Image partition methods can segment pedestrian images into several local regions, which is helpful to extract fine-grained features. However, simple partition often leads to outliers in advanced parts. Therefore, many methods based on key point estimation or semantic segmentation are proposed, which leads to a significant increase in computational cost. Feature descriptors provide distinctive information around key points, which is helpful to enhance feature matching, but it will lead to a large amount of computation consumption. Through this classification, readers cannot only have a systematic understanding of Re-ID, but also have a certain understanding of the current popular research tools. Also, we give the evaluation of some typical models on a set of benchmark datasets and compare them in detail. By discussing their advantages and limitations, we provide a reference for future work. In addition, we briefly introduce some datasets and training tricks.

1.1 Comparison with other surveys

For person Re-ID, state-of-the-art methods have been proposed, and many researchers have surveyed and evaluated them [2, 73, 104, 115, 122]. In [115], research methods are divided into seven categories, the distance metric-based model, the verification model, the identification model, the video-based model, the part-based model, the data augmentation-based model and other methods. In [104], research methods are divided into three branches, feature extraction techniques, metric learning techniques and feature categories. Although these papers have a comprehensive review of the depth model, their structure is not logical. In our paper, we divide the deep models into three categories from coarse to fine according to the scale of feature learning: global-based methods, part-based methods and multiple granularities-based methods. For each classification, we conduct a secondary classification according to the popular research tools, and systematically summarize the state-of-the-art Re-ID methods according to the widely used research tools, such as attention mechanism, feature descriptors, attribute mining, ect. Compared with these papers, our paper is more compact and readable.

Paper [122] is divided into two main categories, open-world person Re-ID and closed-world person Re-ID, and each category is further classified according to the research methods. For example, closed-world Re-ID is divided into feature representation learning methods and deep metric learning methods, etc. Paper [73] mainly summarizes the specific challenges faced by Re-ID, such as mixed dataset feature extraction, location of insufficient labeled data, inadequate hardware resources, etc. Paper [2] divides the Re-ID model into three categories: video based person Re-ID, image based person Re-ID and image to video person Re-ID. The above papers have different classification methods for re ID model, but none of them pay attention to the research tools. In fact, research tools have a great impact on the accuracy of person Re-ID, and play a very important role in the network structure of a deep model. Based on a new classification perspective, that is, the classification based on research tools, we summarize the current popular research tools according to the feature learning scale from coarse to fine.

2 Global-based methods

In general, global-based person Re-ID methods tend to learn features from the whole image to get a feature vector and use this vector to retrieve the gallery images [93, 134]. The initial convolution neural network usually extracts features based on the whole image, thus the global-based person Re-ID methods often do not need too complex a network structure. Global based methods are often suitable for situations with prominent foreground and simple background, but they are poor for scenes with serious occlusion or complex environment, ect. In this section, we introduce the attention-driven methods, the image generation methods and the attribute mining methods for global-based person Re-ID. The attention-driven method introduced in this section is applied to the whole image.

2.1 Attention-driven methods

Attention-driven person Re-ID is an important branch of this field. By simulating the human visual system, attention-driven methods force Re-ID models to pay more attention to the discriminative regions, which can significantly enhance performance of models. Generally, attention models are plug and play, which can be applied to a variety of deep learning methods. However, generation of attention maps will produce a certain amount of computation, which makes the attention-driven methods inapplicable to models requiring low computational consumption. Liu et al. [65] present an end-to-end Comparative Attention Network (CAN). The method they proposed simulates the recognition process of the human visual system by learning a comparison model from the original human images and repeatedly locates the discriminative parts of these images through a set of glimpses. Si et al. [90] propose a novel framework called DuATM, in which the Dual Attention module is composed of an attention layer and a transform layer. Previous methods based on the attention mechanism usually only consider the single intra-sequence attention of features selected from feature sets or sequences, or a single intra-sequence feature refinement for the feature sets or sequences, but do not combine the two, resulting in insufficient to overcome the visual ambiguity in real scenes. Different from these methods, with the dual attention mechanism to simultaneously perform the inter-sequence feature pair alignment and the intra-sequence feature refinement, the DuATM is more robust in the application of real scenes.

On the basis of the bottom-up mechanism, Jian et al. [43] devise a visual-attention-aware mechanism. With an enhanced wavelet-based salient-patch detector, the visually significant patches are captured. Considering human sensitivity to the directional features, a directional-patch detection mechanism is introduced to learn directional patches. The two patches mentioned above are aggregated into the final preferential patches and used for salient-object detection. Similarly, considering the stimulation of orientation information to human visual system, Jian et al. [37] design an innovative computational framework. On the foundation of sparsity criterion and discrete wavelet frame transform, they propose a perceptual directional patch detector to locate salient objects. Meanwhile, a principal local color contrast (PLCC) method is constructed to highlight the salient objects from clutter background with low calculating amount. Considering the challenge of Long-Term Cloth-Changing (LTCC) Re-ID, Qian et al. [80] publish another LTCC dataset and an innovative framework for LTCC Re-ID. The dataset they proposed is the first dataset designed specifically for LTCC Re-ID, containing 17,138 images of 152 identities captured by more than two cameras. The Cloth-Elimination Shape-Distillation (CESD) module they provided in the framework is applied to learn identity related features from cloth appearance features. A Shape Embedding (SE) module is introduced to capture biological pose features via encoded human semantic information.

Although the attention mechanism-based methods have been found efficacious for person Re-ID, some researchers believe that the features learned by these methods are usually relevant and lack diversity. Therefore, Chen et al. [11] devise an Attentive but Diverse Network, which is called ABD-Net to integrate the diversity regularizations and the attention mechanisms simultaneously in a unified network. In ABD-Net, a pair of complementary attention modules are introduced, which respectively focus on position awareness and channel aggregation. Besides, in order to enforce the diversity, they propose a new orthogonality regularizer term. The orthogonal regularizer on the feature space can reduce the feature correlation which is directly conducive to matching, while the orthogonal regularizer on weight can improve the learning ability and promote the filter diversity. The orthogonal regularizer in space can reduce the feature correlation directly beneficial to matching, while the weighted orthogonal regularizer can promote filter diversity and improve the learning ability. Based on the previous work [45], Jian et al. [46] propose a weighted centroid calculation method, which calculates the centroid of salient objects as the prior map and learns the directional information of salient objects. The background features are learned through sparse dictionary to distinguish background and foreground, and suppress the influence of background noise. Simultaneously, the color contrast features are learned and aggregated with the directional information and the background information to form the ultimate saliency map.

However, Chen et al. [9] put forward that many attention-based tasks only consider first-order or rough attention design, and the extracted information is rough and not rich enough. Thus, they propose a higher-order attention mechanism named HOA. Considering the problem of person Re-ID as a zero-shot learning task, they introduce a high order polynomial predictor to model the high-order and complex relationship between visual regions, thus the richness of attention can be enhanced. Zhang et al. [132] present a global-based attention mechanism that both considers the spatial relation-aware attention and the channel relation-aware attention. Considering that the previous methods generally learn attention through local convolutions but ignoring the information mining from the scale of global structure, they propose to learn the attention of each feature node globally by exploring the relationship between features from a global view. To achieve this goal, they propose a relation-aware global attention module, which represents the global range relations and obtains the attention through two convolution layers compactly. Jian et al. [44] devise the first underwater saliency detection method based on the Quaternionic Distance Based Weber Descriptor (QDWD). QDWD independent of image scenes is learned, and the patch distribution of the underwater images is calculate by principal components analysis (PCA) to obtain pattern distinctness. Furthermore, local contrast is used to suppress the background noise and strengthen the salient region.

Bao et al. [5] propose that the major methods of person Re-ID generally focus on the relationship between labels and individual images, ignoring the global mutual information existing in the whole sample set. Thus, they design a Masked Graph Attention Network(MGAN), the core of which is an innovative masked attention mechanism for node updating. The MGAN runs on a complete graph consisting of the extracted features, where under the guidance of label information, nodes can focus on the features of other nodes directionally in the form of a mask matrix. Zhang et al. [127] design an Attention-Aware Scoring Learning (AASL) method, which consists of a score learning head, a Channel Attention Grid (CAG), and a Spatial Attention Grid (SAG). The SAG generates spatial masks, which are fused with the original feature maps to enhance salient regions and suppress irrelevant information. The CAG models the interdependence between the convolution channels to enhance the representation ability of different samples. After the completion of forwarding propagation, the scoring learning head calculates the scores of the CAG and the SAG performance and generates a robust feedback signal to optimize the attention modules. He et al. [25] construct a novel framework named TransRe-ID, which is the first work to utilize a pure transformer for Re-ID task. First, the input image is encoded into a series of patches. Then, a Side Information Embedding module is designed to encodes side information such as viewpoint and camera by learnable embeddings. In the last layer, a Jigsaw Patch Module (JPM) is devised to rearrange patches by shuffle and shift operations, then regroup them for local feature learning.

2.2 Image generation methods

Image generation methods can generate images with different poses, foreground, background, etc., which are suitable for the situation of insufficient training samples and cloth-changing Re-ID, ect. However, there is still a certain gap between images generated by current methods and real images, and these methods are often accompanied by complex network structure and considerable computational consumption. In 2017, Ma et al. [72] proposed a Pose Guided Person Generation Network (PGPGN), which contains two stages in the network. The first stage is a pose integration stage, fed by conditional person image and a target pose to generate a coarse fake image to capture the global structure of the human body under the target pose. With a pose estimator generating the coordinates of key points of the body parts, the model can learn the approximate human body poses without extra annotation of poses. Then with a generator under a U-Net-like architecture, a rough image with basic color and pose close to the target image is generated. In order to mitigate the effect of background changes, they present a new pose mask loss to give the human body more weight than the background. The second stage is an image refinement stage, which aims at refining the results gained at the first stage via adversarial training and generates more accurate images with a variant of Deep Convolutional GAN. The generator at this stage is similar to the one at the first stage, between them the difference is that the second one takes the generated fake image and the original image as its inputs while generates an appearance difference map, and the fully-connected layer of the second one is removed from the U-Net.

Based on the previous work, Ma et al. [71] propose another pose-guided image generation framework named PG^2 in 2018. Similarly, the PG^2 framework is composed of two stages. In the first stage, a multi-branch reconstruction network is introduced to disentangle and encode the background, foreground and pose information into the embedded features, and then these factors are combined to recombine the input image itself. In the second stage, for each factor, three corresponding mapping functions are learned respectively to map the Gaussian noise to the learned embedded feature space. In this manner, pedestrian images with different postures and clothing are generated pertinently. Siarohin et al. [91] argue that, although U-Net-based architectures are commonly used for pose-guided person image generation, the skip connections of the U-Net are badly designed for large spatial deformations because of the misalignment between the local information in the images. Thus, they propose the deformable skip connections to handle the misalignment and translate the local information from the encoder to the decoder under the specific pose differences.

In spite of this, Zhu et al. [151] raise an opinion that images with different poses from different perspectives may have different appearances, which leads to the fact that even with strong deep neural network learning ability, the methods mentioned above cannot produce robust results. To deal with this problem, they provide a progressive pose-guided generative adversarial network, which introduces a new cascaded Pose-Attentional Transfer Blocks (PATBs). Inside each PATB there is an attention mechanism, which infers regions of interest-based on the human pose. When a person's posture and image are recorded, the attention mechanism allows for a better selection of image regions for transfer, guiding the block outputs the generated pose representation and images. In a pose-attentional manner, the model can make better use of the appearance and pose features, thus it transforms the present pose into the target pose more robustly. Balakrishnan et al. [4] design a novel generative adversarial network, which decomposes the human image generation process into a background generation task and a foreground generation task and then combines them to form the final image.

However, in these methods, the generation pipeline and the discriminative Re-ID learning stage are relatively independent of each other. Different from these, Zheng et al. [143] propose a joint learning framework to connect the data generation and the Re-ID learning in an end-to-end way. A self-identity generation module and a cross-identity generation module are introduced to synthesize images of high quality, and a discriminative Re-ID learning module is embedded in the generative module through a shared appearance encoder. Zhang et al. [128] construct a novel pipeline called UnrealPerson. It is completely trained on synthesized data, and outperforms the model trained on real and annotated datasets for the first time. Different from the previous image generation methods, the Unrealperson first generates a series of diversified virtual 3D scenes. Then a random number of pedestrians are generated and put into the generated virtual scene. These pedestrians will move in predefined paths, and their appearance is configurable. Finally, the virtual cameras in the generated scenes capture these images. This method ensures the diversity and authenticity of the generated images, and helps the Re-ID task to achieve better performance. Chen et al. [10] propose to combine a contrastive learning module and a Generative Adversarial Network into a joint training framework. Based on this, they design a mesh based view generator. First, in the unsupervised Re-ID scene, the pedestrian image is disentangled into structural features and identity features. Then the 3D meshes are estimated from the unlabeled training images and rotated to simulate new body structures. Estimated meshes can retain body shape, which makes the generated image retain more visual information. A view-invariant loss is designed for the contrastive module, to lessen the intra-class variation between generated images and real images. The proposed framework does not need labeled source datasets, thus it is more efficient and flexible.

Considering that few people wear the same clothes for a long time in real scenes, Yu et al. [124] design an unsupervised apparel-simulation GAN (AS-GAN) to generate pedestrian images with changed clothes. For the AS-GAN, a pre-trained pixel2pixel model [36] is introduced to learn the cloth mask. The cloth code is captured from a randomly selected image by an auto-encoder network in the same dataset for synthesizing a new image. Ma et al. [71] propose to generate synthetic human images by decomposing the input into three weakly related factors, namely pose, background and foreground. The three factors are learned and encoded into embedding features by a multi-branched reconstruction network. In order to map the Gaussian noise to the embedded feature space respectively for each factor, three corresponding mapping functions learned in an adversarial way are proposed. Tang et al. [101] propose to synthesize clothes-changing images with self-attention mechanism. After extraction of the pose, background and foreground features, a self-attention module is adopted in the foreground encoder to convolve the learned foreground feature maps. The feature vectors obtained by Gaussian noise sampling the foreground features are regarded as the adversarial targets and utilized to generate images with changed foreground features.

2.3 Attribute mining methods

Pedestrian attribute features, such as height, age, and hairstyle, contain high-level semantic information, which has an important impact on person Re-ID. Pedestrian attribute features can be understood as the structured description of pedestrians [118]. They are less sensitive to changes in illumination and viewpoints, and have higher robustness. If an attribute is shared by multiple individuals, learning of this attribute can use the samples of different individuals sharing this attribute for training. Therefore, attribute mining methods can effectively overcome the problem of insufficient category samples. However, because only a few target classes may respond to some attributes, attribute mining methods may face a serious problem of unbalanced category distribution. Layne et al. [51] propose to learn attributes for person Re-ID through a novel data-driven method. A bottom-up attribute ontology is constructed automatically, and the associated representation is learned through large-scale mining of the content on online photo sharing websites. Ontology is automatically captured by clustering comment data and photo tags. A large number of detectors are trained via these clusters, which results in abundant visually detectable attributes. The method proposed by Su et al. [92] has a similar structure to the method proposed by Layne et al. [51] After global pooling, each attribute is assigned a fully connected layer to learn attribute features. Su et al. propose a Low Rank Attribute Embedding, which enhances the capacity of the learned classifiers by applying a low rank embedding to incorporate the complementary attributes into the Re-ID model.

Li et al. [54] design a multi-task learning network called Pedestrian Re-identification network (ARNet) on the basis of attribute mining and reasoning. Different from the above methods [51, 92], the ARNet applies a dual-pooling method to simultaneously learn features containing more location information for attribute identification and features containing the whole image information for person Re-ID. Concurrently, a new attention mechanism is designed to realize the two-dimensional attribute mining of channel and space through the combination of channel and space attention. Luo et al. [22] argue that due to the notable differences between the task of attribute learning and the task of Re-ID, it is likely to be ineffective to directly combine the loss functions of the two. Be aware of this problem, they build an Attribute-identity Feature Fusion Network (AFFNet) to jointly learn identity and attribute on both feature level and loss level. Attributes and identities are learned in two parallel branches and aggregated into the final feature representation. Shi et al. [89] devise an Attribute Mining and Reasoning (AMR) method to mine the value of appearance attributes more deeply. For more accurately localization, they design an Attribute Localization Ensemble (ALE) module. The ALE generates localization scores at each channel and each spatial location with a voting mechanism and multiple localization heads via a weakly-supervised manner. An Attribute Reasoning (AR) module is applied to achieve a more comprehensive person descriptions by aggregating global appearance features and attribute features.

Zhang et al. [126] propose that most of the existing methods utilize attribute information by introducing auxiliary tasks, which may lead to large human consumption for attribute annotations and noisy identity attribute information due to mistaken annotations. Thus, they devise an Attribute Attentional Block (AAB), which generates attention maps by combining attribute and global attention. With an embedded Attribute Selection Module (ASM), the AAB can drop the noisy attributes by reinforcement learning. Quan et al. [81] argue that attention should be paid to the suppression of background information. Aware at this, they construct a new framework combining person discrimination network based on visual attention mechanism and the attribute-identity discrimination network. A multi-instance-learning based image saliency detection is introduced before the network to simultaneously suppress background interference and highlight pedestrians in the foreground. Li et al. [61] design an activation guided identity and attribute classification (AGIAC) framework. The backbone of the AGIAC is divided into four branches for the combination of local features and global features. A branch-guided identity and attribute classification (BIAC) module is introduced to associate each attribute with its reciprocal part, relying on related branches. A mutex local activation (MLA) module is integrated to provide mutually exclusive activation regions for each branch, promoting each branch to collect diversified information. Taking advantage of the attribute prediction method of models trained in a binary classification way which specially designed for hashing, Jin et al. [48] design an attribute based fast retrieval (AFR). An attribute-guided attention block (AAB) is constructed in the AFR to enhance the global feature representation via attribute information.

The performance of some classical methods mentioned in this section is shown in Table 1. Taking mAP and Rank-1 (R-1) as evaluation indicators, we enumerate the top four methods of mAP on Market1501 and Duke-MTMC under each classification, which are arranged in ascending order of mAP.

3 Part-based methods

Part-based person Re-ID methods focus on local images to extract fine-grained and discriminative features. These methods perform well in partial or occluded person Re-ID tasks, but may ignore some global discrimnative information. In this section, we introduce three methods that are widely used in part-based person Re-ID: image partition methods, feature descriptor-based methods, and attention-driven methods. Different from the previous section, the attention-driven methods introduced in this section work on local images.

3.1 Image partition methods

Image partition methods learn features in local regions by partitioning input images or feature maps, which are widely used for fine-grained feature learning. Uniform partition

Table 1 Performance of some global-based methods

Class	Method	Market1501		Method	Duke	
		mAP	R-1		mAP	R-1
Attention-driven	ABD-Net [11]	88.28	95.60	TransReID [25]	82.10	90.70
	RGA-SC [132]	88.40	95.80	AASL [127]	78.94	89.86
	TransReID [25]	88.50	95.20	ABD-Net [11]	78.59	89.00
	AASL [127]	89.56	96.21	MHN [9]	77.20	86.60
Image Generation	GCL [10]	75.40	90.50	FD-GAN [21]	64.50	80.00
	FD-GAN [21]	77.70	90.50	GCL [10]	67.60	81.90
	UnrealPerson [128]	84.70	94.00	UnrealPerson [128]	74.20	86.80
	DG-Net [143]	86.00	94.80	DG-Net [143]	74.80	86.60
Attribute mining	ARNet [54]	87.02	-	ARNet [54]	77.11	-
	AMR [89]	87.68	94.86	AMR [89]	77.24	86.71
	AAB [126]	88.6	96.10	AFR [48]	78.60	88.40
	AFR [48]	90.2	96.60	AAB [126]	80.40	89.90

methods are plug-and-play and have simple structure, but these methods tend to lose the correlation between body parts. Considering this problem, some methods based on key point estimation or semantic segmentation are proposed. However, these methods increase the complexity and computational consumption of models. Some widely-used image segmentation methods are shown in Fig. 5. In 2014, Yi et al. [123] first proposed a simple image partition method for deep Re-ID models. Inspired by the Siamese neural network, they devise a DML framework that divides the input image pair into three overlapped parts and forms them into new image pairs correspondingly, then inputs them into three corresponding siamese convolutional neural networks and calculates the similarity. The PCB model proposed by sun et al. [97] is a very classic image partition method.

In fact, it is a novel part-based convolutional baseline, which performs simple uniform segmentation on the convolution layer to learn local features, and assembles the learned local features into a convolutional descriptor. Based on the PCB [97] model, Chung et al. [15] present an Improved Part-alignment Feature Network (IPAF). They introduce the part identifying learning to fully utilize the features extracted from several body regions, and design a novel part alignment strategy, which normalizes the input image to an accurate and complete human instance profile. Moreover, in the training process, each training image is flipped horizontally as a data enhancement strategy.

However, although the method of image segmentation is simple, it is easy to cause outliers in adjacent parts. To solve this problem, Huang et al. [28] propose a Part Aligned Pooling(PAP) and a Part Segmentation (PS) constraint on the basis of PCB [97] to optimize image partition and enhance alignment. Firstly, the method proposed improves the PCB [97] model by applying a pose estimation to divide body parts under the assistant of key points; secondly, the part segmentation constraint is applied to the feature mapping to further enhance model generalization. The method that Yi et al. [123] provided is based on the hypothesis that the poses of the human body are similar to the spatial distribution of the human body in the bounding box. However, this is often not the case in real scenes. Therefore, Zhao et al. [136] design a party-aligned human representation to solve the problem of body parts misalignment. Inspired by attention-driven deep models, the human body part estimation scheme they proposed is a deep neural network that models the three steps together, learning by minimizing the triplet loss function without body part labeling information. Different from the methods of dividing the image box in space, this method chooses to divide the human body into aligned parts by a part net, thus it can reduce the influence of various human spatial distribution and human posture changes in the bounding box.

Under the scenario of partial person Re-ID, there are some problems such as increasing spatial misalignment and noises from occluded areas. In response to this problem, Sun et al. [95] propose a VPM model, which perceives the features in the visible region through self-supervision learning. The visibility awareness allows VPM to extract local features and focus on the shared visible region of two images for comparison, thus benefits the partial Re-ID. Kalayeh et al. [49] argue that the common methods of image partition tend to simply extract the representation from the horizontal strips, which are loosely related to human body parts. To address this question, they design a SPReID framework, employing human semantic parsing to take advantage of local visual cues for person Re-ID. Under the framework, the input RGB images are converted into an activation tensor by convolution backbone. Simultaneously, the probability maps related to diverse semantic regions of the human body are generated by a human semantic parsing branch. Zhang et al. [131] design a novel framework based on densely semantically aligned person Re-ID. By estimating the dense semantics of human images, they design a series of densely semantically aligned partial images (DSAP images), in which the same spatial positions share the same semantics between diverse images. They design a dual-stream framework, which is comprised of a Densely Semantically-aligned Guiding Stream (DSAG Stream) and a Main Full Image Stream (MF Stream). The DSAG stream with DSAP image as input is used as a regulator,

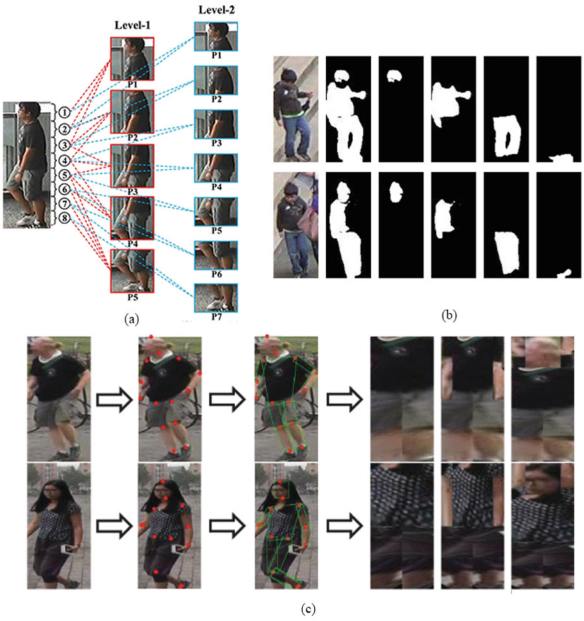


Fig. 5 Some image partition methods. **a** Horizontal stripes based method [30]; **b** Semantic segmentation based method [82]; **c** Key points based method [140]

which instructs the MF stream to extract semantically aligned features intensively from the initial image.

Huang et al. [29] propose that many existing methods are based on the partition of human parts with horizontal stripes or semantic segmentation, separately. Thus, they design a Multi-scale Discriminative network with Region Segmentation (MDRS), which integrates horizontal stripe partition, semantic segmentation, and multi-scale discriminative feature learning in a unified network. In this framework, the number of stripes increases with scale to obtain more fine-grained information. The task of human part segmentation will make the multi-scale feature maps more discriminative, which can improve the performance of the Re-ID module through the shared feature maps. Considering that the semantic segmentation is uncertain, Tu et al. [103] propose a semantic alignment model based on an entropy-based mask. The model aligns human semantic features according to the confidence score and the visible score defined by entropy. It dynamically aligns the confidence and common semantic human regions without extra computational load.

3.2 Feature descriptor based methods

Feature descriptors are a set of vectorized descriptions of images, which only contain the key information of images. Due to the uniqueness of generated vectors, taking feature descriptors as basis of matching can enhance the robustness of models to rotation, scaling, viewpoint change, illumination change and so on. However, some feature descriptors are designed manually, which could lead to considerable labor consumption. In addition, feature descriptor based methods may face the problems of long feature matching time and high computational consumption. In some shallow networks, such methods may be difficult to achieve accurate matching and local feature descriptor extraction under complex conditions. Zhu et al. [149] argue that many existing person Re-ID databases are not large enough to train deep models. Thus, they present a body symmetry and part-locality-guided direct non-parametric deep feature enhancement (DNDFE). The model consists of two non-parametric layers: the local normalization layer and the symmetrical average pool layer. The symmetrical average pool is introduced to enhance the feature maps in the light of the perpendicular bisector. The local normalization layer is designed to make sure that the feature maps of different regions distribute in the same range. Matsukawa et al. [74] propose a novel feature descriptor on the basis of the hierarchical distribution of pixel features, which describes the local regions in the images by hierarchical Gaussian distribution. In this model, local patches are extracted intensively in a region, and the region is regarded as a group of local patches. Firstly, the region is modeled as a group of multiple Gaussian distributions, each of which indicates the appearance of a local patch, which is dubbed patch Gaussian. Then, another Gaussian distribution is used to describe the characteristics of patch Gaussian, which is called region Gaussian. Finally, the parameters of region Gaussian are introduced as feature vectors to represent the image region.

Based on this work [74], Matsukawa et al. [75] provide a novel meta-descriptor based on hierarchical Gaussian distribution. Different from the previous work, in both steps, they embed the parameters of the Gaussian distribution into a point of the Symmetric Positive Definite (SPD) matrix manifold, and design a normalization method of feature norm to alleviate the bias trend on the SPD matrix descriptor. Gou et al. [23] propose that also Gaussian descriptors can achieve state-of-the-art performance, it is not the case if the feature distribution isn¡t Gaussian. Therefore, they design a novel descriptor on the basis of the on-manifold mean of the moment matrix. With the empirical moment matrix to combine higher-order moments, and the on-manifold mean to gather features around the horizontal stripes, it can approximate the distribution of non-Gaussian and more complex distributed pixel features in medium-sized local patches.

Inspired by the previous works [84, 86], Satta et al. [85] devise a dissimilarity representation that can be introduced to the appearance-based, multiple instance descriptors and provide compact descriptors with low matching time. This type of descriptor is called the multiple component dissimilarity descriptors (MCD). In this manner, the human body is divided into an upper body part and a lower body part, and each part is represented by a set of components. The clustering algorithm is introduced to generate multiple prototypes for each part. The component set representing each body part of an individual is compared with the prototype related to the part, and the value obtained forms a different vector part. Pala et al. [78] propose that the Re-ID precision of a clothing appearance descriptor can be enhanced by fusing the RGB-D sensor with the anthropometric measures extracted from the depth data under unconstrained conditions. Based on the MCD descriptor [85], they design a fusion method on the basis of feature-level fusion.

Patruno et al. [79] provide another model, which works on the RGB-D data. Firstly, it introduces a point cloud pre-processing to clean up the outliers and noise. Then a human Skeleton Standard Posture (SSP) is defined by utilizing aligned skeleton joints. The SSP is a novel representation of the human skeleton under fixed postures, which can be used to divide the point cloud into different generated grids and define distinguishing features. Finally, the color-based descriptors are obtained by investigating each divided region. Wu et al. [116] devise a deep multiplicative integration gating function and generate joint descriptors for human matching. The framework they proposed is divided into the upper layer and the lower layer. The lower layer consists of two-stream CNNs, and the outputs of the final convolution are combined by a multiplicative integration gate at each position. The upper layer corresponds to the stacked four-directional recurrent layers and obtains the spatial relationship through lateral connections. The resulting joint features can be applied to similarity measurement by reducing cross-view misalignment.

Tan et al. [100] propose a Consecutive Batch DropBlock Network (CBDB-Net), which can generate efficient pedestrian descriptor from incomplete feature maps. The Consecutive Batch DropBlock Modules (CBDBMs) they designed divide the feature maps uniformly, and continuously attach each patch to the feature map independently from top to bottom on the convolutional layer. The gained incomplete feature maps are utilized in the training stage to obtain robust descriptor. Considering local feature misalignment, Huang et al. [32] provide a Validity aggregation and multi-scale feature extraction network (VMSFEN). The VMSFEN consists of three branches, each branch learn features independently. The learned features are aggregated into the final feature descriptor through a method combining the local feature cross-alignment strategy and the validity aggregation strategy. Sun et al. [96] design an enhanced PCB, based on the PCB mentioned above. After consolidation, the PCB can adapt to different partition strategies, and the RPP is strengthened to gain attentive feature descriptors. Experiments show that PCB based on uniform partition (PCB-U) achieves higher accuracy. Wan et al. [105] insist that both local region representation and local region discovery should be considered for the attention mechanism. Thus, they provide a constrained attention module, consisting of an iterative concentration process and a multi-scale attention module. Under the constraint of the iterative concentration process, the multi-scale attention module generates concentrated local parts based on middle-level feature maps. A statistical-positional-relational (SPR) descriptor is designed for the description of local regions.

3.3 Attention-driven methods

Considering that the pose information is not fully utilized, Xu et al. [119] devise an Attention-Aware Compositional Network (AACN). It is composed of two branches, the Attention-aware Feature Composition (AFC) branch and the Pose guided Part Attention (PPA) branch. The PPA branch is designed to evaluate the visibility score and attention map for each predefined part of the human body. Under the given guidance of the visibility score and attention map from the PPA branch, the AFC performs part feature alignment and weighted fusion. Similarly, considering the importance of the pose information, Gao et al. [20] design a Pose-guided Visible Part Matching (PVPM) model. The PVPM is composed of two main parts: a pose-guided visibility predictor (PVP) and a pose-guided part attention(PGA) mechanism. The PVP predicts the visibility of image parts in a self-supervised manner, while the PGA fuses the pose-guided attention maps with the appearance features.

Another pose-guided method, which is called Pose-Guided Feature Alignment (PGFA) [76], is proposed by Miao et al.. In the feature construction stage, the PGFA uses human landmarks to generate attention maps to indicate whether a concrete body part is occluded or not, and guide the model to focus on the non-occluded parts. In the matching stage, the global feature is divided into several parts, and the pose landmarks are used to indicate which part of the features belong to the target person. Use only visible regions for retrieval. Xu et al. [120] present an Attentional Part-based CNN (AP-CNN) method, which integrates attention mechanism into local feature learning. It divides the feature map into several horizontal stripes and applies an attention mechanism to each stripe. For the attention mechanism, they design a free-parameter attention method with a skip layer connection to maximize the complementary information between the feature recognition and the attention selection. Liang et al. [63] design a Related Attention Network (RAN) for person Re-ID. Firstly they provide a key point-based data pre-processing algorithm to align the human images into a standard template. Secondly, they propose a novel attention mechanism, concentrating on the human body parts under a pixel level correlation.

Zhang et al. [125] propose a Local Heterogeneous Features (LHF) method, to learn local features from three aspects: local compact features, local relative features, and local discriminative features. To learn the local discriminative features, the attention maps are divided into three horizontal parts and the classification operation is performed. As for the local compact features and the local relative features, the attention maps are divided into two parts, and the center loss, as well as the triplet loss, are utilized to learn them respectively. Zhou et al. [148] design a foreground attentive neural network (FANN) to learn the discriminative features from the foreground of an input image. The FANN focuses on the foreground bypassing each input image through a network of encoders and decoders. The gained feature maps are averagely segmented and deeply learned in a body part sub-network. Then, the generated feature maps are fused by a feature fusion sub-network. Finally, the ultimate feature vectors are normalized to the unit spherical space, and the symmetric triplet loss layer learning is followed. Zhang et al. [133] devise the Heterogeneous Local Graph Attention Networks (HLGAT), which can simultaneously model the intra-local relation within single image and the inter-local relation among different images. The images are fed into CNN to capture feature maps, and local features are extracted by uniform partition strategy and global max pooling operation. These local features are used to generate a complete local graph, and an attention mechanism is used to aggregate the local features during the learning process of intra-local relation and inter-local relation. An attention regularization loss is constructed to constrain the attention weights for the inter-local relation, while the contextual information is introduced into the attention weights to inject structure information for the intra-local relation.

Huo et al. [33] construct an Attentive Part-aware Networks (APN) designed for partial person Re-ID. Considering the misalignment of partial person Re-ID, they cut out the effective part of the whole image, and proposed a Cropping Type Consistency (CTC) loss for the classification of cropping types. A Block Attention Mechanism (BAM) is incorporated to strengthen the consistent partial features learned with the help of CTC. Lin et al. [64] propose to randomly crop the images from a full-body dataset, e.g. the Market1501, to mimic query images of partial Re-ID. An image rescaler is adopted to generate distortion-free images in a self-supervised manner from query inputs. A bodymap-to-appearance (B2A) co-attention is designed to capture visible body regions, and perform partial image matching based on body parsing maps generated by a pre-trained human parsing model.

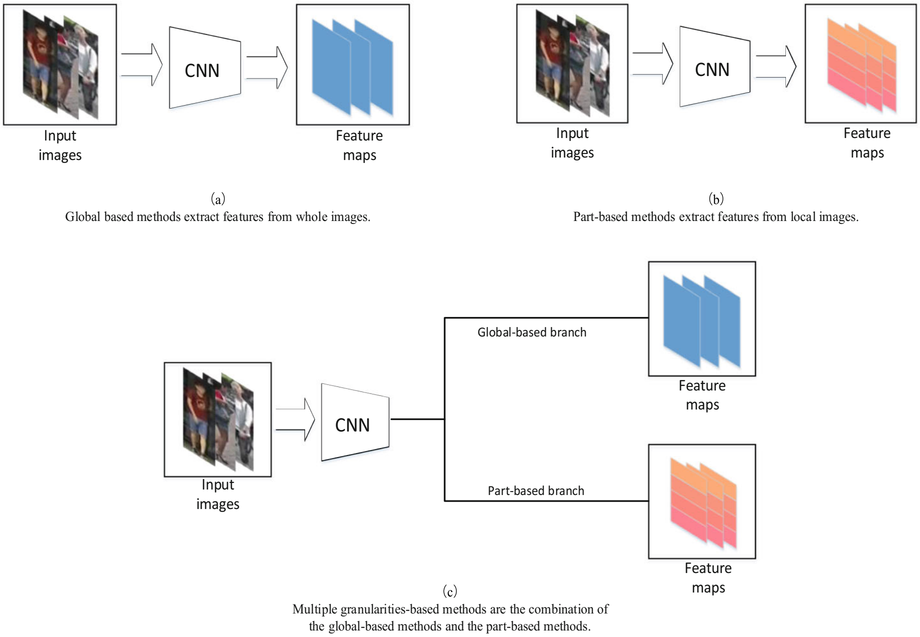
The performance of some classical methods mentioned in this section is shown in Table 2. Taking mAP and Rank-1 (R-1) as evaluation indicators, we enumerate the top four methods of mAP on Market1501 and Duke-MTMC under each classification, which are arranged in ascending order of mAP.

4 Multiple granularities-based methods

The global-based methods are based on the overall situation, but they tend to ignore some crucial details; although the part-based methods can capture these ignored fine-grained features, they are less holistic. The multiple granularities-based methods are the combination of the global-based methods and the part-based methods and have the advantages of both, as shown in Fig. 6. In this section, we introduce two widely-used methods for multiple granularities-based person Re-ID: image partition methods and attention-driven methods.

Table 2 Performance of some part-based methods

Class	Method	Market1501		Mehod	Duke	
		mAP	R-1		mAP	R-1
Image Partition	IPAF [15]	85.96	94.30	DSAP [131]	74.30	86.20
	Esa-Reid [103]	86.30	94.50	Esa-Reid [103]	77.60	87.70
	DSAP [131]	87.60	95.70	MDRS [29]	79.40	89.40
	MDRS [29]	87.60	95.80	IPAF [15]	84.70	89.84
Feature Descriptor	PCB-U [96]	81.60	93.80	PCB-U [96]	71.50	84.50
	CSPR-Net [105]	84.80	-	CSPR-Net [105]	71.90	-
	CBDB-Net [100]	85.20	94.40	CBDB-Net [100]	73.90	87.30
	VMSFEN [32]	88.60	96.30	VMSFEN [32]	80.30	89.90
Attention-driven	LHF [125]	86.10	94.90	FANN [148]	70.20	85.20
	AP-CNN [120]	82.70	94.00	AP-CNN [120]	75.00	87.30
	AACN [119]	82.96	88.69	LHF [125]	77.60	87.10
	HLGAT [133]	93.40	97.50	HLGAT [133]	87.30	92.70



(a)
Global based methods extract features from whole images.

(b)
Part-based methods extract features from local images.

(c)
Multiple granularities-based methods are the combination of the global-based methods and the part-based methods.

Fig. 6 Multiple granularities-based methods have the advantages of both the global-based methods and the part-based methods

4.1 Image partition methods

The Multiple Granularity Network (MGN) [108] that Wang et al. proposed achieves very good results. It is a multi-branch architecture that consists of one global feature learning branch and two local feature learning branches. The local branches divide the image evenly into several stripes and perform the same operation as the global branch on each part to obtain fine-grained feature representation. In order to learn both global and local robust feature representations of human body parts, Li et al. [55] devise a Multi-Scale Contex-tAware Network (MSCAN). Each convolutional layer of the network adopts several dilated convolution kernels with diverse receptive fields, then the feature maps gained are concatenated as the output of the current layer. Considering that the human body is nonrigid, the spatial transformation network is used to localize latent pedestrian parts instead of using the predefined rigid region. In order to adapt the network to the pedestrian partial localization task, they devise three new constraints on the learned transformation parameters. Therefore, the possibility of pedestrian image sample misalignment is reduced.

Although pose estimation can better enhance the alignment, it needs a large number of labeled data and considerable GPU memory to obtain human pose heatmaps. Therefore, Luo et al. devise [70] a Dynamically Matching Local Information (DMLI) method, which can dynamically align local information with no additional supervision. Under the assumption that for two images of the same person, the features of their corresponding parts of the body share higher similarity (smaller distance), after obtaining the processed features of two pedestrian images, the feature distance of the two images is calculated and the distance matrix is constructed, then the shortest path planning is used to automatically find similar slices for alignment. Fu et al. [19] propose a Horizontal Pyramid Matching (HPM) model.

The HPM divides the deep feature maps into several horizontal parts of multiple scales for a Horizontal Pyramid Pooling operation, and the resulting feature representations of each part are then utilized to learn multi-scale information independently. The max and the average pooling features in each partition are fused to learn a person-specific discriminative representation in a global-local way. Zhang et al. [129] devise a semantic-aware occlusion-robust network (SORN) for partial and occluded person Re-ID. The SORN consists of three branches, a global branch, a local branch, and a semantic branch. A spatial-patch contrastive loss (SPC) is proposed for the global branch to capture occlusion-robust global features. A foreground-background mask is generated in the semantic branch to indicate the visible body parts. Inspired by PCB [97], the feature maps are evenly divided in the local branch and fed to the average pooling to learn discriminative features.

Considering the problem that in the complex real scene, the existing detection model often produces imprecise bounding boxes, which has a negative impact on the performance of Re-ID models, Zheng et al. [139] propose another coarse-to-fine pyramid model to ease the dependence on the bounding boxes. The pyramid is able to capture discriminative information at different scales, and segment the feature map layer by layer from coarse to fine. With a dynamic training scheme introduced, the pyramid model can integrate not only global and local information but also the progressive clues between them. Zhao et al. [137] argue to learn person Re-ID by a series of designed human body parts. By grid partition, they divide the whole image into the vertical, the horizontal, the partial, and the occlusive parts. A CNN model is trained for each part separately in the training stage. By cascading the features of corresponding parts and the whole image, the extra part information is utilized to evaluate the similarity between the gallery images and the probe image.

Inspired by the PCB method [97], Li et al. [60] propose a Multi-Scale Feature Fusion Network. In the local branch of the proposed network, the average pooling is introduced to generate a globally pooled feature map. The resulting feature map is divided into six horizontal stripes, and these stripes are trained by six cross-entropy loss functions separately. Considering that the RPP module in the PCB [97] model makes the network more complicated and the global information is ignored, they suggest replacing the RPP module with a global branch and combine it with the local branch. Wei et al. [113] design a Global-Local-Alignment Descriptor (GLAD), as well as a retrieval framework. The GLAD divides the input image into three parts on the basis of four detected key points. A four-stream deep neural network is designed to learn robust representation on both the global and local levels.

Similarly based on a key point estimation [7] to align the input image, Wang et al. devise [110] a two-module framework, which learns the weighted local body part and global features jointly. A global pooling is directly applied to capture global features. Observing that different body parts are of different discrimination, they propose to divide the image based on the structure of the human body unequally, for the feature extraction of local features. The extracted local and global features are then fused by the method provided in [142]. Zhao et al. [135] present a Spindle Net, which is also based on the key point to divide body regions. The network extracts semantic features from the generated body regions separately, and the resulting features are merged with a competitive scheme. The features with different semantic levels are merged during different stages, rather than directly concatenated together. The method that Rodolfo et al. [82] proposed is the first work that combines saliency and semantic parsing in a unified framework. The framework proposed is composed of two branches. One is called S-ReID sub-network, which focuses on obtaining global saliency features, while the other named SP-ReID sub-network focuses on obtaining local semantic parsing features.

Considering the challenge of Long-Term Cloth-Changing (LTCC) Re-ID, Huang et al. [31] publish a large-scale LTCC dataset named Celebrities-reID. The Celebrities-reID contains 10,842 pedestrian images of 590 identities, and nobody in this dataset wears the same clothing twice. A two-step fine-tuning strategy on human body parts (2SF-BPart) is proposed as the benchmark approach of the Celebrities-reID. Considering that different body parts of people wearing different clothes play different roles in matching, the 2SF-BPart partitions pedestrian images via the pose estimation to learn each body part respectively, and simultaneously takes into account the feature learning of the whole image with the assist of the two-stream IDE (2S-IDE) neural network [144]. Wan et al. [106] provide a small real dataset and a large-scale synthetic dataset for Cloth-Changing Re-ID. Meanwhile, they propse to introduce a local face feature extractor to detect faces, and all the cut face images are resized to 50 × 50 pixel size. Two 512 dimensional feature vectors learned from the face feature extractor and the holistic feature extractor are concatenated via a weighted sum.

4.2 Attention-driven methods

Li et al. [58] propose a new attention network called HA-CNN to jointly learn both the soft pixel-level and the hard region-level attention and feature representations for person Re-ID. In order to learn different kinds of attention effectively from the shared Re-ID feature representation, they designed a lightweight harmonious attention module, which can learn in a multitasking and end-to-end manner. In addition, in order to further improve the compatibility between feature description and attention selection, a cross-attention interactive learning scheme is introduced. Tay et al. [102] devise a new architecture called Attribute Attention Network (AAN), based on a novel attention mechanism called attribute attention. This network divides person Re-ID into three branches: the Global Feature Network (GFN), the Attribute Feature Network (AFN), and the Part Feature Network (PFN). The outputs of these three tasks are combined using homoscedastic uncertainty learning to predict the person identification. It is worth mentioning that attribute attention is a very novel concept. The Attribute Feature Network captures the key attribute information, such as hair and clothing color. This information is firstly utilized to performs classification on personal attributes, and the output of this stage is used to generate a class activation map for each attribute. Finally, the Attribute Feature Network combines the class activation map generated by the selected attribute classes into a feature map and submits it to the AAM classifier for learning.

Based on the Strong Baseline [69], Tan et al. [99] propose a novel CNN model which is simple but efficient for person Re-ID, named MSBA. Between stages of the backbone, they directly introduce the squeeze-and-excitation modules [27] as attention modules to strengthen channel attention. Moreover, in addition to introducing two branches from the second and third stages to generate multi-scale features to improve the generalization ability of the model, they also introduce an extra local branch at the fourth stage to enhance the representation of the body features. However, Chen et al. [12] offer an opinion that those methods have the limitation that although they pay close attention to the most salient features, they often ignore some relatively less salient features, such as backpacks, shoes, and so on. In fact, these less significant features often have a very important impact on the accuracy of pedestrian recognition. In response to this problem, Chen et al. present a Salience-guided Cascaded Suppression Network (SCSN) to enable the model to mine features with different salience at the same time. It introduces two new components, a feature aggregation module, and a salient feature extraction unit. With a residual dual attention module and a non-local fusion block, the feature aggregation module is able to restrain the salient features gained in the previous stage and learn other potential salient features adaptively. For the salient feature extraction unit, the feature map is divided into several stripes, then each stripe is explored by a convolutional layer, a BN layer, and a ReLU layer for mining fine-grained information as well as dimension reduction. Finally, a salience selector is introduced to learn the salience-sensitive weights and judge whether inhibition is needed accordingly.

Inspired by HOA [9], Cai et al. [130] design a novel multi-scale body-part mask-guided attention network, which combines both spatial attention and channel attention. The network consists of two attention modules. One is under the guidance of the whole-body mask, while the other is divided into three parts, guided by bottom-body mask, upper-body mask, and whole-body mask, respectively. Liu et al. [66] propose a HydraPlus-Net, which fully utilizes local and global information with multi-level feature fusion. The multi-directional attention module that they proposed extracts features from the attentive regions of multiple layers. The generated attention maps are fed to different feature layers multi-directionally, and the global and local features are integrated into a final feature vector. Considering the significance of pose information, Gong et al. [22] design a two-stream network, which combines the pose estimation and the attention mechanism. They suggest fusing the high-level and the middle-level features and correlate global features by the self-attention mechanism to discriminate the view-invariant features from different semantic levels. Meanwhile, they introduce a pose estimation stream to guide the self-attention module for taking the edge information of the human body into consideration. A bilinear pooling is utilized to aggregate the captured features into final features. Sun et al. [94] design a Local to Global with Multi-scale Attention Network (LGMANet) consisting of two branches. The local to global branch generates three feature maps of different spatial scales by a pooling generation. Then the feature maps are segmented into several horizontal bins. During this stage, the local information of different spatial scales can be fused with the final global information by different segmentation methods. The multi-scale attention branch is introduced to extract the contextual dependencies from different layers.

Inspired by the Convolutional Block Attention Module (CBAM) [114] and the Squeeze-and-Excitation Network (SENet) [27], Zhong et al. [145] propose a Part based Attention Model (PAM), which consists of both the spatial attention block and the channel attention block. The input image is evenly divided into several parts, and each part is fed into the PAM to refine the spatial and channel feature. The refined feature maps of the global and local body parts are fused into the global and local feature representation, each feature representation is trained by identity classification loss. Yang et al. [121] design an attention-driven multi-branch network, in which each branch introduces an intra-attention network to obtain the discriminative parts in the body-part or whole-body images. Similar to the PAM [145], the attention modules provided in this paper consider both the channel-wise attention and the spatial-wise attention. They propose to fuse the inter-attention module and the intra-attention module in an end-to-end manner. The inter-attention module is designed to adaptively fuse local and global features, while the intra-attention network independently learns features from the precisely aligned local or global images. Li et al. [62] design a Part-Aware Transformer (PAT), which is the first work to deal with occluded person Re-ID via a transformer encoder-decoder architecture integrated in a unified deep framework. It is composed of a part prototype based transformer decoder and a pixel context based transformer encoder. A self-attention mechanism is introduced into the pixel context based transformer encoder to learn the context information of the full image. The part prototype based transformer decoder is designed to capture discriminative body parts with only identity labels.

In the part prototype based transformer decoder, a self-attention mechanism is adopted to further integrate the local context of body parts, while a cross-attention layer is incorporate to learn foreground part features. Based on ViT [17], Sharma et al. [88] propose a Locally Aware Transformer (LA-Transformer). Considering that although the output of the visual transformer is mainly a global classification token, it also generates additional information about the local parts, they adopt a Parts-based Convolution Baseline (PCB [97])-inspired strategy to aggregate globally enhanced local classification tokens. With a blockwise fine-tuning incorporated, the model's performance on Re-ID task further improves. Zhu et al. [150] propose that the lack of a clear alignment mechanism weakens transformer's ability for person Re-ID task. Thus, they construct an Auto-Aligned Transformer (AAformer). First, the input images are divided into patches of fixed size, each patch is embedded linearly, and the position embeddings are added. Then, the learnable vectors of part tokens and class token are introduced to generate part and global feature representations separately. Similar to paper [88], the proposed framework is also based on ViT [17]. However, the Multi-head Self-Attention in ViT [17] is displaced by the Multi-head Auto-Alignment proposed in this paper to achieve part alignment.

The performance of some classical methods mentioned in this section is shown in Table 3. Taking mAP and Rank-1 (R-1) as evaluation indicators, we enumerate the top four methods of mAP on Market1501 and Duke-MTMC under each classification, which are arranged in ascending order of mAP.

5 Benchmark datasets

For the research of person Re-ID, reliable data set is very important foundation. The changes of illumination and pose in the data sets, as well as their complex background and various occlusion, play an important role in the training of person Re-ID models. So far, numerous data sets have been proposed for person Re-ID. In this section, we mainly introduce the widely used data sets, like VIPeR [24], GRID [68], PRID [26], CUHK01-03 [56, 57, 59], Market1501 [141], DukeMTMC-reID [83], MSMT17 [112] and a newly published data set CrowdHuman [87], as shown in Table 4.

VIPeR [24] Viper is an earlier published data set for person Re-ID, which is characterized by the diversity of illumination and viewpoint, is quite challenging because each individual

Table 3 Performance of some multiple granularities-based methods

Class	Method	Market1501		Method	Duke	
		mAP	R-1		mAP	R-1
Image Partition	MFFN [60]	85.90	95.00	MFFN [60]	76.00	87.30
	MGN [108]	86.90	95.70	MGN [108]	78.40	88.70
	Pyramid [139]	88.20	95.70	Pyramid [139]	79.00	89.00
	SSP-REID [82]	90.80	93.70	SSP-REID [82]	83.70	86.40
Attention-driven	PAT [62]	88.00	95.40	PAT [62]	78.20	88.80
	MSBA [99]	89.00	95.80	SCSN [12]	79.00	90.10
	SCSN [12]	88.50	95.70	MSBA [99]	79.70	90.30
	LA-Transformer [88]	94.46	98.27	AAformer [150]	80.00	90.10

Table 4 State-of-the-art data sets

Dataset	Time	ID	Image	Cam.	Label	Eval.	Size
VIPeR	2007	632	1264	2	Hand	CMC	128*48
GRID	2009	1025	1275	8	Hand	CMC	Vary
PRID	2011	934	24541	2	Hand	CMC	128*64
CUHK01	2012	971	3884	2	Hand	CMC	160*60
CUHK02	2013	1816	7264	10	Hand	CMC	160*60
CUHK03	2014	1467	13164	10	Hand\DPM	CMC	Vary
Market1501	2015	1501	33217	6	Hand\DPM	C\M	128*64
DukeMTMC-reid	2017	1812	36441	8	Hand	C\M	Vary
MSMT17	2018	4101	126441	15	Faster RCNN	C\M	Vary
CrowdHuman	2018	339565	24370	Vary	Hand	R\A	Vary

has only two images. It contains 1264 images taken from two different cameras, including 632 identities. These images are hand-labeled and adjusted to 128 × 48 pixels.

GRID [68] The QMUL underground Re-ID data set, which is called GRID, is collected by 8 uncrossed and uncalibrated cameras at a subway station. Among these cameras, two platforms are equipped with three cameras each, while the other two monitor a connection junction point far away from the platforms. The quality of these images collected in this data set is quite poor, and the activities in the scene are very complex, which makes the data set considerably challenging. The whole data set consists of 1275 images of 1025 identities.

PRID [26] The PRID data set consists of two versions, one is the single-shot version, and the other is the multi-shot version. In the single-shot version, each identity only has one image which is randomly selected. While in the multi-shot version, each identity has at least five images collected from per camera. Images in this data set are collected from two different static surveillance cameras, and there are great differences in background, pose, illumination, etc. This data set contains 24541 images of 934 identities.

CUHK01-03 [56, 57, 59] In 2012, Li et al. presented their first data set CUHK01 [59], which contains 971 identities and per identity has two images collected in two non-intersecting cameras. One year later, they proposed the second data set CUHK02 [56]. Compared with the earlier CUHK01, there are up to ten camera views in the CUHK02 data set. These camera views are divided into five pairs, each pair has 971, 107, 306, 239, and 193 identities respectively, and per identity has two images collected in each camera. On the basis of cuhk01 and cuhk02 data sets, in 2014 they proposed the CUHK03 data set [57], which is the first data set that is huge enough to train DNN models. The whole data set contains 13, 164 images of 1, 360 identities. With six surveillance cameras fixed, each identity is captured by two non-intersecting camera views, averaging 4.8 images per view. It not only provides the pedestrian images that are manually cropped but also provides the samples which are detected by the pedestrian detector automatically. In addition, in this dataset, the samples collected from the five pairs of camera views are admixed together, and thus form a complex cross-view transformation. Therefore, compared with the previous data sets, this data set is more complex and closer to the real scene.

Market1501 [141] Market1501 is a data set collected on the campus of Tsinghua University in the summer. With five high definition cameras and a low definition camera used, a total of 1501 pedestrian images were collected. In the training set, there are 12,936 images of 751 identities; while in the testing set, there are 19,732 images of 750 identities. Different from the CUHK03 data set using the DPM detector, except for false positive boxes, the Market1501 data set also provide false alarms.

DukeMTMC-reid [83] Collected by eight outdoor cameras, the DukeMTMC-reID data set contains 36,441 images of 1,812 identities. Among them, 702 identities are randomly selected to form the training set, and the others to form the test set. The trajectories are manually annotated with an interface to mark key points and associate identities between cameras. Thus, with the key points automatically interpolated, each identity has ground-plane world coordinates and single-frame bounding boxes of all the cameras in which it is collected.

MSMT17 [112] The MSMT17 data set is collected by 12 indoor cameras and three outdoor cameras, consisting of 126,441 images of 4,101 identities. For capturing the initial video, four days with diverse climate conditions are randomly selected during a month, Every morning, noon, and afternoon, and three hours of video were taken respectively in the morning, noon, and afternoon per day. Therefore, the images collected in this data set have high diversity, and the illumination and pose variation of these images are complex. Thus, it is closer to the actual scene.

CrowdHuman [87] This is the first data set specifically aimed at the human detection task of crowd problems, and the average number of people in the image is as high as 22.6. This data set is considerably huge, contains 15,000 images for training, 4370 images for verification, and 5000 images for testing. There are up to 339,565 identities in all the images, with various kinds of occlusion. More than this, it provides three kinds of annotation boxes, a full bounding box for each instance, a head bounding box for the head area, and a visible bounding box for the unoccluded parts.

6 Training tricks

Effective training tricks have positive effects on the performance of the model, and some of them are introduced in many papers, such as [69, 70, 99]. In this section, we will give a brief overview of some popular training tricks.

Random Erasing [147] Zhong et al. propose a data augmentation method to train the convolutional neural network named Random Erasing. In the training stage, the Random Erasing selects a rectangle region randomly with a certain probability, to erase pixels of the selected region with random values. Thus, training samples with multifarious occlusion are generated, and the robustness of the model to occlusion can be enhanced. The area proportion and aspect ratio of erasing areas can be controlled by parameters.

K-Reciprocal Encoding [146] Zhong et al. design a k-reciprocal encoding mechanism for the re-ranking of Re-ID results. They provide a hypothesis that if the gallery image is similar to the probe in the k-reciprocal nearest neighbor, it seems more likely to be the real match. The weighted k-reciprocal neighbor set is firstly encoded into a vector and forms the k-reciprocal feature. Subsequently, the Jaccard distance of the two images is calculated according to their k-reciprocal features. A local query expansion mechanism is developed to further achieve more robust k-reciprocal features. As the final distance, the weighted sum of the original distance and the Jaccard distance will be calculated.

BNNeck [70] The combination of triplet loss and ID loss is a widely used method to train Re-ID model [3, 14, 16]. However, considering that the optimization objectives of the two loss functions might not be consistent, it is possible that one loss is oscillating while the other is decreasing. To address the problem, Luo et al. propose a BNNeck mechanism. BNNeck simply applies a batch normalization (BN) layer between features and the FC layers. The features before BN layer are introduced to train the triplet loss, while the normalized features after BN are used to train the ID loss. The features are Gaussian distribution around the surface of the hypersphere. In this way, the ID loss is easier to converge, and the constraints on ID loss are reduced. Thus, the triple loss is easier to converge at the same time.

Label smoothing [98] Considering that there are usually a small number of false labels in the learning samples, which may influence the performance of the model, Szegedy et al. present a Label Smoothing mechanism, which assumes that there may be errors in the label during training to prevent over-fitting. Suppose there are a smoothing parameter and a distribution on the label u (k), which is independent of the training example X. For the training example with a ground-truth label y, the process can be divided into two stages. Firstly, set the label k as the ground-truth label k = y. Secondly, with a certain probability, k is replaced by the sample extracted from the distribution u(k). They suggest that using the prior distribution on the labels as u(k).

7 Future directions and current limitations

For the person Re-ID tasks, complex models tend to achieve higher accuracy. Therefore, researchers have designed many multi-branch models and deep network frameworks. The complex network structure causes huge calculation consumption, which suffers from challenges in training and parameter adjustment, and makes the research cost for person Re-ID increases greatly. Therefore, despite the good performance in experiments, some of the existing person Re-ID models cannot be used in the real sense. It should be an important research direction that how to use the lightweight network structure and less calculation consumption to achieve the ideal Re-ID accuracy.

Moreover, despite the publication of several large-scale datasets, the training samples are still far from enough. Although the training samples can be expanded by image generation, there are still considerable differences between true images and generated images. Current image generation methods are difficult to simulate the complex occlusion and illumination changes in the real scene, as well as the differences between different monitoring devices. Meanwhile, most of the existing person Re-ID methods acquiesce in the existence of bounding boxes, that is to say, person detection and person Re-ID are carried out separately in the great majority of the existing methods. However, in real application scenarios, person detection, and person Re-ID often need to be carried out simultaneously. Therefore, how to integrate both tasks into a unified framework, and meanwhile consider the negative impact of inaccurate bounding boxes on the performance of models, is still a problem to be solved.

Considering that it is almost impossible for people to wear the same clothes for a long time in real scenes, Long-Term Cloth-Changing (LTCC) Re-ID has important research significance. At present, the methods for cloth-changing Re-ID are mainly divided into two categories: learning cloth-invariant feature representation and generating clothing-changed samples through image generation methods to train models. The research on cloth-changing Re-ID is insufficient and should be paid attention to. Simultaneously, in real monitoring environment, serious occlusion often occurs, and the collected pedestrian images are often partial. Therefore, partial Re-ID and occluded Re-ID should also be further explored.

In addition, with the popularity of UAVs, more and more crowded places tend to utilize UAVs for object detection and monitoring. However, the research on pedestrian recognition in aerial images is relatively poor. To solve this problem, we will carry out further research in the future.

8 Conclusion

Due to the limitations of training samples, illumination variation, camera sensors differences, pose variation, and other inherent limitations, person Re-ID is still a challenging task and has received extensive attention. In recent years, person Re-ID has been widely studied. In this paper, we give a comprehensive review of some current methods for person Re-ID. Firstly, we discuss the challenges faced by person Re-ID and its significance. Secondly, we briefly review the current methods based on deep learning. We divide these deep learning-based methods into three categories: global-based methods, part-based methods, and multiple granularity-based methods. Based on each classification, a secondary classification is made according to the learning method of models. Thirdly, we provide some descriptions of popular person Re-ID datasets and summarize several common training tricks. In addition, we summarize the performance of some mentioned methods on the three most popular datasets in recent years. We hope this paper can provide a reference for future works.

Declarations

Conflict of Interests Author Minjie Liu, Author Jiaqi Zhao, Author Yong Zhou, Author Hancheng Zhu, Author Rui Yao and Author Ying Chen declare that they have no conflict of interest.

Author Jiaqi Zhao has received research grants No. 61806206 from the National Natural Science Foundation of China and No. BK20180639 from the Natural Science Foundation of Jiangsu Province.

Author Yong Zhou has received research grants No. BK20201346 from the Natural Science Foundation of Jiangsu Province and No. 2015-DZXX-010 from the Six Talent Peaks Project in Jiangsu Province.

Author Hancheng Zhu has received research grants No. 62101555 from the National Natural Science Foundation of China and No. BK20210488 from the Natural Science Foundation of Jiangsu Province.

Author Rui Yao has received research grants No. 62172417 from the National Natural Science Foundation of China and No. 2018-XYDXX-044 from the Six Talent Peaks Project in Jiangsu Province.

References

1. Abbas Q, Ibrahim MEA, Jaffar MA (2019) A comprehensive review of recent advances on deep vision systems. Artif Intell Rev 52(1):39–76
2. Almasawa MO, Elrefaei LA, Moria K (2019) A survey on deep learning-based person re-identification systems. IEEE Access 7:175228–175247. https://doi.org/10.1109/ACCESS.2019.2957336

3. Bai X, Yang M, Huang T, Dou Z, Yu R, Xu Y (2020) Deep-person: learning discriminative deep features for person re-identification. Pattern Recogn 98:107036. https://doi.org/10.1016/j.patcog.2019.107036
4. Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J (2018) Synthesizing images of humans in unseen poses. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 8340–8348
5. Bao L, Ma B, Chang H, Chen X (2019) Masked graph attention network for person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1496–1505
6. Becerra-Riera F, Morales-González A, Méndez-Vázquez H (2019) A survey on facial soft biometrics for video surveillance and forensic applications. Artif Intell Rev 52(2):1155–1187
7. Cao Z, Simon T, Wei S, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1302–1310
8. Chang Y-S, Wang M-Y, He L, Lu W, Su H, Gao N, Yang X-A (2018) Joint deep semantic embedding and metric learning for person re-identification. Pattern Recogn Lett 130. https://doi.org/10.1016/j.patrec.2018.08.011
9. Chen B, Deng W, Hu J (2019) Mixed high-order attention network for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 371–381
10. Chen H, Wang Y, Lagadec B, Dantcheva A (2021) Joint generative and contrastive learning for unsupervised person re-identification
11. Chen T, Ding S, Xie J, Yuan Y, Chen W, Yang Y, Ren Z, Wang Z (2019) Abd-net: Attentive but diverse person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 8350–8360
12. Chen X, Fu C, Zhao Y, Zheng F, Song J, Ji R, Yang Y (2020) Salience-guided cascaded suppression network for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3297–3307
13. Chen Y, Zhu X, Gong S (2019) Instance-guided context rendering for cross-domain person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 232–242
14. Cheng D, Gong Y, Shi W, Zhang S (2018) Person re-identification by the asymmetric triplet and identification loss function. Multimed Tools Appl 77(3):3533–3550. https://doi.org/10.1007/s11042-017-5182-z
15. Chung S, Xue Y, Chien S, Chan R (2019) Improved part-aligned deep features learning for person re-identification. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 1–8
16. Dai P, Ji R, Wang H, Wu Q, Huang Y (2018) Cross-modality person re-identification with generative adversarial training. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18. AAAI Press, pp 677–683
17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale
18. Franco A, Oliveira L (2016) A coarse-to-fine deep learning for person re-identification. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 1–7
19. Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang T (2018) Horizontal pyramid matching for person re-identification. Proc AAAI Conf Artif Intelli 33 https://doi.org/10.1609/aaai.v33i01.33018295
20. Gao S, Wang J, Lu H, Liu Z (2020) Pose-guided visible part matching for occluded person reid. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 11741–11749
21. Ge Y, Li Z, Zhao H, Yin G, Yi S, Wang X, Li H (2018) Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, pp 1230–1241
22. Gong X, Zhu S (2019) Person re-identification based on two-stream network with attention and pose features. IEEE Access 7:131374–131382. https://doi.org/10.1109/ACCESS.2019.2935116
23. Gou M, Camps O, Sznaier M (2017) mom: Mean of moments feature for person re-identification. In: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), pp 1294–1303
24. Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro
25. He S, Luo H, Wang P, Wang F, Li H, Jiang W (2021) Transreid: Transformer-based object re-identification
26. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Heyden A, Kahl F (eds) Image Analysis, pp 91–102

27. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
28. Huang H, Yang W, Chen X, Zhao X, Huang K, Lin J, Huang G, Du D (2018) Eanet: Enhancing alignment for cross-domain person re-identification
29. Huang J, Liu B, Fu L (2020) Joint multi-scale discrimination and region segmentation for person re-id. Pattern Recogn Lett 138 https://doi.org/10.1016/j.patrec.2020.08.022
30. Huang Y, Sheng H, Zheng Y, Xiong Z (2017) Deepdiff: Learning deep difference features on human body parts for person re-identification. Neurocomputing 241:191–203. https://doi.org/10.1016/j.neucom.2017.02.055
31. Huang Y, Wu Q, Xu J, Zhong Y (2019) Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp 1–8
32. Huang Z, Qin W, Luo F, Guan T, Xie F, Han S, Sun D (2021) Combination of validity aggregation and multi-scale feature for person re-identification. Journal of Ambient Intelligence and Humanized Computing
33. Huo L, Song C, Liu Z, Zhang Z (2021) Attentive part-aware networks for partial person re- identification. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp 3652–3659
34. Islam K (2020) Person search: New paradigm of person re-identification: A survey and outlook of recent works. Image Vis Comput:103970
35. Islam K (2020) Person search: New paradigm of person re-identification: A survey and outlook of recent works. Image Vis Comput 101:103970. https://doi.org/10.1016/j.imavis.2020.103970
36. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5967–5976
37. Jian M, Zhang W, Yu H, Cui C, Nie X, Zhang H, Yin Y (2018) Saliency detection based on directional patches extraction and principal local color contrast. J Vis Commun Image Represent 57:1–11
38. Jian M, Cui C, Nie X, Zhang H, Nie L, Yin Y (2019) Multi-view face hallucination using svd and a mapping model. Inf Sci 488:181–189. https://doi.org/10.1016/j.ins.2019.03.026
39. Jian M, Lam K-M (2014) Face-image retrieval based on singular values and potential-field representation. Signal Process 100:9–15. https://doi.org/10.1016/j.sigpro.2014.01.004
40. Jian M, Lam K-M (2015) Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. IEEE Trans Circ Syst Video Technol 25(11):1761–1772. https://doi.org/10.1109/TCSVT.2015.2400772
41. Jian M, Lam K-M, Dong J (2013) A novel face-hallucination scheme based on singular value decomposition. Pattern Recogn 46(11):3091–3102. https://doi.org/10.1016/j.patcog.2013.03.020
42. Jian M, Lam K-M, Dong J (2014) Facial-feature detection and localization based on a hierarchical scheme. Inf Sci 262:1–14. https://doi.org/10.1016/j.ins.2013.12.001
43. Jian M, Lam K-M, Dong J, Shen L (2015) Visual-patch-attention-aware saliency detection. IEEE Trans Cybern 45(8):1575–1586. https://doi.org/10.1109/TCYB.2014.2356200
44. Jian M, Qi Q, Dong J, Yin Y, Lam K-M (2018) Integrating qdwd with pattern distinctness and local contrast for underwater saliency detection. J Vis Commun Image Represent 53:31–41. https://doi.org/10.1016/j.jvcir.2018.03.008
45. Jian M, Wang J, Yu H (2019) Visual saliency detection via background features and object-location cues. In: 2019 25th International Conference on Automation and Computing (ICAC), pp 1–4
46. Jian M, Wang J, Yu H, Wang G, Meng X, Yang L, Dong J, Yin Y (2021) Visual saliency detection by integrating spatial position prior of object with background cues. Expert Syst Appl 168:114219. https://doi.org/10.1016/j.eswa.2020.114219
47. Jiao S, Wang J, Pan Z, Hu G, Zou J, Zeng M (2019) Multi-layer joint classification-metric deep learning for top view image person re-identification. In: 2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE), pp 47–50
48. Jin H, Lai S, Zhao G, Qian X (2021) Hashing person re-id with self-distilling smooth relaxation. Neurocomputing 455:111–124. https://doi.org//10.1016/j.neucom.2021.05.059
49. Kalayeh MM, Basaran E, Gökmen M, Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1062–1071
50. Lavi B, Ullah I, Fatan M, Rocha A (2020) Survey on reliable deep learning-based person re-identification models: Are we there yet?
51. Layne R, Hospedales T, Gong S (2014) Re-id: Hunting attributes in the wild. In: British Machine Vision Conference

52. Lei J, Niu L, Fu H, Peng B, Huang Q, Hou C (2019) Person re-identification by semantic region representation and topology constraint. IEEE Trans Circ Syst Video Technol 29(8):2453–2466. https://doi.org/10.1109/TCSVT.2018.2866260
53. Leng Q, Ye M, Tian Q (2020) A survey of open-world person re-identification. IEEE Trans Circ Syst Video Technol 30(4):1092–1108. https://doi.org/10.1109/TCSVT.2019.2898940
54. Li C, Yang X, Yin K, Chang Y, Wang Z, Yin G (2021) Pedestrian re-identification based on attribute mining and reasoning. IET Image Process 15(11):2399–2411. https://doi.org/10.1049/ipr2.12225
55. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification
56. Li W, Wang X (2013) Locally aligned feature transforms across views. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp 3594–3601
57. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 152–159
58. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2285–2294
59. Li W, Zhao R, Wang X (2013) Human reidentification with transferred metric learning. In: Lee KM, Matsushita Y, Rehg JM, Hu Z (eds) Computer Vision – ACCV 2012, pp 31–44
60. Li Y, Wang X, Zhu Z, Huang X, Li P, Qi G, Rong Y (2020) A novel person re-id method based on multi-scale feature fusion. In: 2020 39th Chinese Control Conference (CCC), pp 7154–7159
61. Li Y, Zhang B, Sun J, Chen H, Zhu J (2021) Person re-identification based on activation guided identity and attribute classification model. Multimed Tools Appl (1)
62. Li Y, He J, Zhang T, Liu X, Zhang Y (2021) Diverse part discovery: Occluded person re-identification with part-aware transformer
63. Liang J, Zeng D, Chen S, Tian Q (2019) Related attention network for person re-identification. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp 366–372
64. Lin C-S, Wang Y-CF (2021) Self-supervised bodymap-to-appearance co-attention for partial person re-identification. In: 2021 IEEE International Conference on Image Processing (ICIP), pp 2299–2303
65. Liu H, Feng J, Qi M, Jiang J, Yan S (2017) End-to-end comparative attention networks for person re-identification. IEEE Trans Image Process 26(7):3492–3506. https://doi.org/10.1109/TIP.2017.2700762
66. Liu X, Zhao H, Tian M, Sheng L, Shao J, Yi S, Yan J, Wang X (2017) Hydraplus-net: Attentive deep features for pedestrian analysis. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 350–359
67. Liu X, Xia Y, Yu H, Dong J, Jian M, Pham TD (2020) Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation. IEEE Trans Neural Syst Rehab Eng 28(10):2325–2332. https://doi.org/10.1109/TNSRE.2020.3021410
68. Loy CC, Xiang T, Gong S (2009) Multi-camera activity correlation analysis. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 1988–1995
69. Luo H, Gu Y, Liao X, Lai S, Jiang W (2019) Bag of tricks and a strong baseline for deep person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 1487–1495
70. Luo H, Jiang W, Zhang X, Fan X, Qian J, Zhang C (2019) Alignedreid++: Dynamically matching local information for person re-identification. Pattern Recogn 94:53–61
71. Ma L, Sun Q, Georgoulis S, Van Gool L, Schiele B, Fritz M (2018) Disentangled person image generation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 99–108
72. Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L (2017) Pose guided person image generation. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, vol 30, pp 406–416
73. Mathur N, Mathur S, Mathur D, Dadheech P (2020) A brief survey of deep learning techniques for person re-identification. In: 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), pp 129–138
74. Matsukawa T, Okabe T, Suzuki E, Sato Y (2016) Hierarchical gaussian descriptor for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1363–1372
75. Matsukawa T, Okabe T, Suzuki E, Sato Y (2020) Hierarchical gaussian descriptors with application to person re-identification. IEEE Trans Pattern Anal Mach Intell 42(9):2179–2194. https://doi.org/10.1109/TPAMI.2019.2914686
76. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-guided feature alignment for occluded person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 542–551
77. Neves J, Narducci F, Barra S, Proença H (2016) Biometric recognition in surveillance scenarios: a survey. Artif Intell Rev 46(4):515–541

78. Pala F, Satta R, Fumera G, Roli F (2016) Multimodal person reidentification using rgb-d cameras. IEEE Trans Circ Syst Video Technol 26(4):788–799. https://doi.org/10.1109/TCSVT.2015.2424056
79. Patruno C, Marani R, Cicirelli G, Stella E, D'Orazio T (2019) People re-identification using skeleton standard posture and color descriptors from rgb-d data. Pattern Recogn 89:77–90
80. Qian X, Wang W, Zhang L, Zhu F, Fu Y, Xiang T, Jiang Y-G, Xue X (2021) Long-term cloth-changing person re-identification. In: Ishikawa H, Liu C-L, Pajdla T, Shi J (eds) Computer Vision – ACCV 2020, pp 71–88
81. Quan H, Feng S, Lang C, Chen B (2020) Improving person re-identification via attribute-identity representation and visual attention mechanism. Multimed Tools Appl 79(11):7259–7278
82. Quispe R (2019) Improved person re-identification based on saliency and semantic parsing with deep neural network models. Image Vis Comput 92:103809
83. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: Hua G, Jégou H (eds) Computer Vision – ECCV 2016 Workshops, pp 17–35
84. Satta R, Fumera G, Roli F (2011) Exploiting dissimilarity representations for person re-identification, vol 7005, pp 275–289
85. Satta R, Fumera G, Roli F (2012) Fast person re-identification based on dissimilarity representations. Pattern Recogn Lett 33(14):1838–1848
86. Satta R, Fumera G, Roli F, Cristani M, Murino V (2011) A multiple component matching framework for person re-identification. In: Image Analysis and Processing – ICIAP 2011
87. Shao S, Zhao Z, Li B, Xiao T, Yu G, Zhang X, Sun J (2018) Crowdhuman: A benchmark for detecting human in a crowd. arXiv:1805.00123
88. Sharma C, Kapil SR, Chapman D (2021) Person re-identification with a locally aware transformer
89. Shi Y, Wei Z, Ling H, Wang Z, Shen J, Li P (2020) Person retrieval in surveillance videos via deep attribute mining and reasoning. IEEE Trans Multimed:1–1. https://doi.org/10.1109/TMM.2020.3042068
90. Si J, Zhang H, Li C, Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5363–5372
91. Siarohin A, Sangineto E, Lathuilière S, Sebe N (2018) Deformable gans for pose-based human image generation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3408–3416
92. Su C, Yang F, Zhang S, Tian Q, Davis LS, Gao W (2015) Multi-task learning with low rank attribute embedding for person re-identification. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 3739–3747
93. SU J, He X, Qing L, Yu Y, Xu S, Peng Y (2019) A new discriminative feature learning for person re-identification using additive angular margin softmax loss. In: 2019 UK/ China Emerging Technologies (UCET), pp 1–4
94. Sun L, Liu J, Zhu Y, Jiang Z (2019) Local to global with multi-scale attention network for person re-identification. In: 2019 IEEE International Conference on Image Processing (ICIP), pp 2254–2258
95. Sun Y, Xu Q, Li Y, Zhang C, Li Y, Sun J (2019) Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification, pp 393–402
96. Sun Y, Zheng L, Li Y, Yang Y, Tian Q, Wang S (2021) Learning part-based convolutional features for person re-identification. IEEE Trans Pattern Anal Mach Intell 43(3):902–917. https://doi.org/10.1109/TPAMI.2019.2938523
97. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018, pp 501–518
98. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision
99. Tan H, Xiao H, Zhang X, Dai B, Lai S, Liu Y, Zhang M (2020) Msba: Multiple scales, branches and attention network with bag of tricks for person re-identification. IEEE Access 8:63632–63642. https://doi.org/10.1109/ACCESS.2020.2984915
100. Tan H, Liu X, Bian Y, Wang H, Yin B (2021) Incomplete descriptor mining with elastic loss for person re-identification. IEEE Trans Circ Syst Video Technol:1–1. https://doi.org/10.1109/TCSVT.2021.3061412
101. Tang C, Guo J (2020) Clothes-changing image generation based on attention for person re-identification. In: 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pp 2009–2013
102. Tay C, Roy S, Yap K (2019) Aanet: Attribute attention network for person re-identifications. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7127–7136
103. Tu C, Zhao Y, Cai L (2020) Esa-reid: Entropy-based semantic feature alignment for person re-id

104. Vidhyalakshmi MK, Poovammal E (2017) A survey on recent approaches in person re-id. In: Dash SS, Vijayakumar K, Panigrahi BK, Das S (eds) Artificial intelligence and evolutionary computations in engineering systems, pp 503–511
105. Wan C, Wu Y, Tian X, Huang J, Hua X-S (2020) Concentrated local part discovery with fine-grained part representation for person re-identification. IEEE Trans Multimed 22(6):1605–1618. https://doi.org/10.1109/TMM.2019.2946486
106. Wan F, Wu Y, Qian X, Chen Y, Fu Y (2020) When person re-identification meets changing clothes. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 3620–3628
107. Wang G, Zhang T, Cheng J, Liu S, Yang Y, Hou Z (2019) Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 3622–3631
108. Wang G, Yuan Y, Chen X, Li J, Zhou X (2018) Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp 274–282
109. Wang K, Wang H, Liu M, Xing X, Han T (2018) Survey on person re-identification based on deep learning. CAAI Trans Intell Technol 3(4):219–227. https://doi.org/10.1049/trit.2018.1001
110. Wang Y, Wang Z, Jia W, He X, Jiang M (2018) Joint learning of body and part representation for person re-identification. IEEE Access 6:44199–44210. https://doi.org/10.1109/ACCESS.2018.2864588
111. Wang Z, Wang Z, Zheng Y, Chuang Y, Satoh S (2019) Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 618–626
112. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 79–88
113. Wei L, Zhang S, Yao H, Gao W, Tian Q (2017) Glad: Global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia, pp 420–428
114. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018, pp 3–19
115. Wu D, Zheng SJ, Zhang XP, Yuan CA, Cheng F, Zhao Y, Lin YJ, Zhao ZQ, Jiang YL, Huang DS (2019) Deep learning-based methods for person re-identification: a comprehensive review. Neurocomputing 337(APR.14):354–371
116. Wu L, Wang Y, Li X, Gao J (2018) What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. Pattern Recogn 76:727–738
117. Wu W, Tao D, Li H, Yang Z, Cheng J (2020) Deep features for person re-identification on metric learning. Pattern Recogn:107424
118. Xiao W, Sz A, Rui YA, Az B, Zhe CD, Jin TA, Bl A (2021) Pedestrian attribute recognition: A survey. Pattern Recogn
119. Xu J, Zhao R, Zhu F, Wang H, Ouyang W (2018) Attention-aware compositional network for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2119–2128
120. Xu Y, Jiang Z, Men A, Pei J, Ju G, Yang B (2019) Attentional part-based network for person re-identification. In: 2019 IEEE Visual Communications and Image Processing (VCIP), pp 1–4
121. Yang F, Yan K, Lu S, Jia H, Xie X, Gao W (2019) Attention driven person re-identification. Pattern Recogn 86:143–155
122. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SCH (2021) Deep learning for person re-identification: A survey and outlook. IEEE Trans Pattern Anal Mach Intell:1–1. https://doi.org/10.1109/TPAMI.2021.3054775
123. Yi D, Lei Z, Liao S, Li SZ (2014) Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition, pp 34–39
124. Yu Z, Zhao Y, Hong B, Jin Z, Huang J, Cai D, He X, Hua X-S (2021) Apparel-invariant feature learning for person re-identification. IEEE Trans Multimed:1–1. https://doi.org/10.1109/TMM.2021.3119133
125. Zhang H, Si T, Zhang Z, Zhang R, Ma H, Liu S (2020) Local heterogeneous features for person re-identification in harsh environments. IEEE Access 8:83685–83692. https://doi.org/10.1109/ACCESS.2020.2991838
126. Zhang J, Niu L, Zhang L (2021) Person re-identification with reinforced attribute attention selection. IEEE Trans Image Process 30:603–616. https://doi.org/10.1109/TIP.2020.3036762

127. Zhang M, Xin M, Gao C, Wang X, Zhang S (2020) Attention-aware scoring learning for person re-identification. Knowl-Based Syst 203:106154. https://doi.org/10.1016/j.knosys.2020.106154

128. Zhang T, Xie L, Wei L, Zhuang Z, Zhang Y, Li B, Tian Q (2021) Unrealperson: An adaptive pipeline towards costless person re-identification

129. Zhang X, Yan Y, Xue J-H, Hua Y, Wang H (2021) Semantic-aware occlusion-robust network for occluded person re-identification. IEEE Trans Circ Syst Video Technol 31(7):2764–2778. https://doi.org/10.1109/TCSVT.2020.3033165

130. Zhang Y, Guo J, Huang Z, Qiu W, Fan H (2019) Multi-scale body-part mask guided attention for person re-identification. MATEC Web Conf 277(2):02025

131. Zhang Z, Lan C, Zeng W, Chen Z (2019) Densely semantically aligned person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 667–676

132. Zhang Z, Lan C, Zeng W, Jin X, Chen Z (2020) Relation-aware global attention for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3183–3192

133. Zhang Z, Zhang H, Liu S (2021) Person re-identification using heterogeneous local graph attention networks

134. Zhao C, Chen B, Chen B (2020) Human parsing with discriminant feature learning for person re-identification. In: ICRSA 2020: 2020 3rd International Conference on Robot Systems and Applications, pp 30–34

135. Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X (2017) Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 907–915

136. Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 3239–3248

137. Zhao Y, Li Y, Wang S (2020) Person re-identification with effectively designed parts. Tsinghua Sci Technol 25(3):415–424. https://doi.org/10.26599/TST.2019.9010031

138. Zhao Y, Lin J, Xuan Q, Xi X (2019) Hpiln: a feature learning framework for cross-modality person re-identification. IET Image Process 13(14):2897–2904. https://doi.org/10.1049/iet-ipr.2019.0699

139. Zheng F, Deng C, Sun X, Jiang X, Guo X, Yu Z, Huang F, Ji R (2019) Pyramidal person re-identification via multi-loss dynamic training. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8506–8514

140. Zheng L, Huang Y, Lu H, Yang Y (2019) Pose-invariant embedding for deep person re-identification. IEEE Trans Image Process 28(9):4500–4509. https://doi.org/10.1109/TIP.2019.2910414

141. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 1116–1124

142. Zheng L, Wang S, Tian L, Fei He, Liu Z, Tian Q (2015) Query-adaptive late fusion for image search and person re-identification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1741–1750

143. Zheng Z, Yang X, Yu Z, Zheng L, Yang Y, Kautz J (2019) Joint discriminative and generative learning for person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2133–2142

144. Zheng Z, Zheng L, Yang Y (2017) A discriminatively learned cnn embedding for person reidentification. ACM Trans Multimed Comput Commun Appl 14(1). https://doi.org/10.1145/3159171

145. Zhong W, Jiang L, Zhang T, Ji J, Xiong H (2020) A part-based attention network for person re-identification. Multimedia Tools and Applications 79. https://doi.org/10.1007/s11042-019-08395-2

146. Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3652–3661

147. Zhong Z, Zheng L, Kang G, Li S, Yang Y (2017) Random erasing data augmentation. Proc AAAI Conf Artif Intell 34. https://doi.org/10.1609/aaai.v34i07.7000

148. Zhou S, Wang J, Meng D, Liang Y, Gong Y, Zheng N (2019) Discriminative feature learning with foreground attention for person re-identification. IEEE Trans Image Process 28(9):4671–4684. https://doi.org/10.1109/TIP.2019.2908065

149. Zhu J, Zeng H, Huang J, Zhu X, Lei Z, Cai C, Zheng L (2020) Body symmetry and part-locality-guided direct nonparametric deep feature enhancement for person reidentification. IEEE Internet Things J 7(3):2053–2065. https://doi.org/10.1109/JIOT.2019.2960549

150. Zhu K, Guo H, Zhang S, Wang Y, Huang G, Qiao H, Liu J, Wang J, Tang M (2021) Aaformer: Auto-aligned transformer for person re-identification

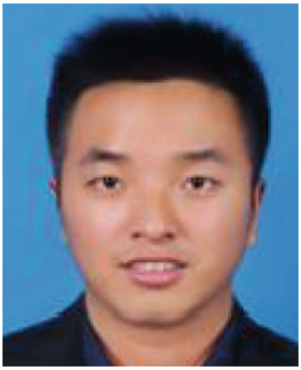
151. Zhu Z, Huang T, Shi B, Yu M, Wang B, Bai X (2019) Progressive pose attention transfer for person image generation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2342–2351
152. Zhuo J, Chen Z, Lai J, Wang G (2018) Occluded person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp 1–6

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Minjie Liu received the B.E. degree in Internet of Things Engineering from Jiangsu University of Science and Technology, Zhenjiang, China, in 2017. She is currently working toward the M.S. degree at the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou. Her research interests include computer vision and person re-identification.

Jiaqi Zhao received the B.E. degrees in intelligence science and technology in 2010, the Ph.D. degree in circuits and systems in 2017 from Xidian University, Xi'an, China. Between 2013-2014, he was an exchange Ph.D. student with the Leiden Institute for Advanced Computer Science (LIACS), University of Leiden, the Netherlands. He is currently with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His current research interests include multiobjective optimization, deep learning and image processing.

Yong Zhou received the Ph.D. degree from the Department of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. He is now a professor in China University of Mining and Technology. His research mainly focuses on data mining, machine learning and artificial intelligence.



Hancheng Zhu received the B.S. degree from the Changzhou Institute of Technology, Changzhou, China, in 2012, and the M.S. and Ph.D. degree from the China University of Mining and Technology, Xuzhou, China, in 2015 and 2020. His research interests include affective computing and image aesthetics assessment.

Rui Yao received the Ph.D. degree in computer science from Northwestern Polytechnical University, Xian, China, in 2013. From September 2011 to September 2012, he was a Visiting Student with the University of Adelaide, Adelaide, SA, Australia. He is currently associate professor with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China. His current research interests include computer vision and machine learning.

Ying Chen received the B.S. degree in software engineering from Shandong University, Weihai, China, in 2012. She is currently working toward the Ph.D. degree at the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou. Her research interests include computer vision, Generative Adversarial Networks, person re-identification and visual quality assessment.