Complex & Intelligent Systems
https://doi.org/10.1007/s40747-023-01229-7

ORIGINAL ARTICLE

A comprehensive review of pedestrian re-identification based on deep learning

Zhaojie Sun[1] · Xuan Wang[1] · Youlei Zhang[2] · Yongchao Song[1] · Jindong Zhao[1] · Jindong Xu[1] · Weiqing Yan[1] · Cuicui Lv[1]

Received: 28 November 2022 / Accepted: 14 August 2023

Abstract
Pedestrian re-identification (re-ID) has gained considerable attention as a challenging research area in smart cities. Its applications span diverse domains, including intelligent transportation, public security, new retail, and the integration of face re-ID technology. The rapid progress in deep learning techniques, coupled with the availability of large-scale pedestrian datasets, has led to remarkable advancements in pedestrian re-ID. In this paper, we begin the study by summarising the key datasets and standard evaluation methodologies for pedestrian re-ID. Second, we look into pedestrian re-ID methods that are based on object re-ID, loss functions, research directions, weakly supervised classification, and various application scenarios. Moreover, we assess and display different re-ID approaches from deep learning perspectives. Finally, several challenges and future directions for pedestrian re-ID development are discussed. By providing a holistic perspective on this topic, this research serves as a valuable resource for researchers and practitioners, enabling further advancements in pedestrian re-ID within smart city environments.

Keywords Pedestrian re-identification · Deep learning · Large-scale datasets · Smart cities

Introduction

The topic of pedestrian re-ID is commonly regarded as a sub-problem of image retrieval. Unfortunately, due to differences in the environment and camera view, good quality or high-resolution (HR) photographs of faces are frequently unavailable in surveillance footage. When face re-ID fails, re-ID becomes an important alternative technology. Re-ID is an approach that utilizes computer vision technology to identify a specific pedestrian in an image or video using non-overlapping cameras [1]. It is an ability to identify the same target pedestrians in different scenes based on characteristics, such as clothing, body type, hairstyle, etc., and is also a technique used for cross-border tracking.

The pedestrian re-ID problem can be divided into two parts: pedestrian re-ID and cross-camera tracking. Given a surveillance video with pedestrian images, obtain images of one pedestrian from different cameras. When matched and found using the appropriate procedures, the queried pedestrian can be described through photos [2, 3], video images, and sequences [4, 5], as well as textual descriptions [6]. Additional description and application methods are also used in many complex circumstances, such as re-ID between depth

✉ Xuan Wang
xuanwang91@ytu.edu.cn

Zhaojie Sun
sunzhaojie@s.ytu.edu.cn

Youlei Zhang
zhangyoulei@wismake.com

Yongchao Song
ycsong@ytu.edu.cn

Jindong Zhao
zhjdong@ytu.edu.cn

Jindong Xu
xujindong@ytu.edu.cn

Weiqing Yan
wqyan@tju.edu.cn

Cuicui Lv
lvcuicui@ytu.edu.cn

[1] School of Computer and Control Engineering, Yantai University, Yantai 264005, Shandong, China

[2] Yantai Huitong Network Technology Co., Yantai 264005, Shandong, China

Published online: 03 October 2023

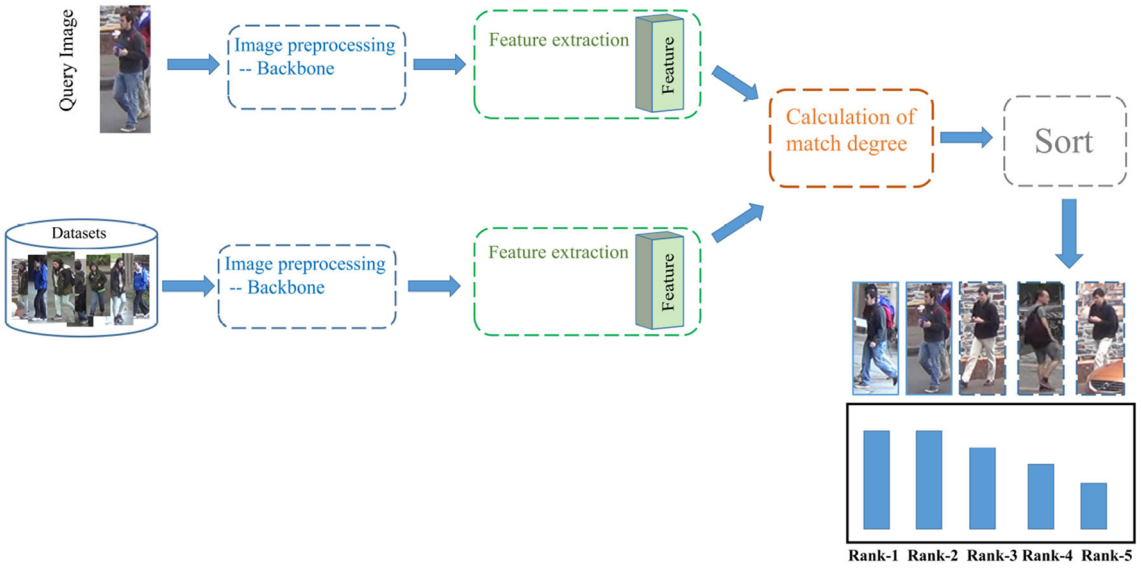

Fig. 1 Pedestrian re-ID system

and RGB images [7, 8], text-to-image re-ID [9, 10], visible low-resolution (LR) re-ID [11], cross-resolution re-ID [12], and so on.

Re-ID reduces the tediousness and inefficiency of pedestrian retrieval work, decreases the restrictions of current camera perspective change, and can be coupled with smart city pedestrian identification and tracking technology. Other fields, such as intelligent security, intelligent pedestrian-finding systems, and unmanned supermarkets of smart super and intelligent robots, offer a wide range of applications with high application and development potential as well as practical usefulness.

The pedestrian re-ID system is depicted in Fig. 1. The algorithm flow of the system is as follows: (1) the original data set is fed into the backbone network to filter pedestrians. (2) The screened pedestrians and the images of pedestrians to be queried are fed into the pre-trained network to extract features. (3) The matching degree is calculated, and the pedestrians with the highest similarity are chosen based on the similarity ranking.

Although current re-ID technology offers more benefits than standard video surveillance, there are certain technical issues and challenges. Because there are variations in different camera views [13], differences between devices as well as interference from noisy background environments [14] are influenced by various movements and reactions of pedestrians in different situations [15, 16] and are prone to many problems: blurred and LR [12], variations in lighting in the environment [17]. Early re-ID research mainly focused on multi-camera tracking, manual feature creation around body shape and structure [18–20], and distance metric learning.

With the continuous development of deep learning, re-ID techniques have made significant progress in detection precision and experimental feasibility [2, 21, 22]. Several re-ID methods proposed by researchers in recent years have improved the accuracy of matching pedestrian in networks and models, with significantly better results than previous re-ID methods. Figure 2 depicts the numerous circumstances that could arise during the re-ID task.

Between 1997 and 2019, re-ID technology underwent significant development. The progression of re-ID technology can be divided into two stages based on time: the stage of manual feature extraction methods before 2014 and the stage of deep learning-based approaches after 2014. The continuous advancement in computer vision has been instrumental in the evolution of re-ID technology. It has shifted from manual feature acquisition to the utilization of deep learning methods, leading to numerous milestone breakthroughs. Figure 3 illustrates key stages in the technological development process.

Compared with previous work [23–27], the differences of this review are as follows: first, we provided a comprehensive overview of key datasets and evaluation methods, laying a solid foundation for pedestrian re-ID research. Second, we focused on studying deep learning-based methods for pedestrian re-ID and their potential applications in smart cities, highlighting their performance advantages. Additionally, we conducted detailed investigations into the performance of relevant models on different datasets and made improvements to the classification methods. Finally, we delved into the challenges and future directions in the field of pedestrian re-ID, offering valuable guidance for researchers and practitioners. In summary, this paper stands out for its comprehensiveness,

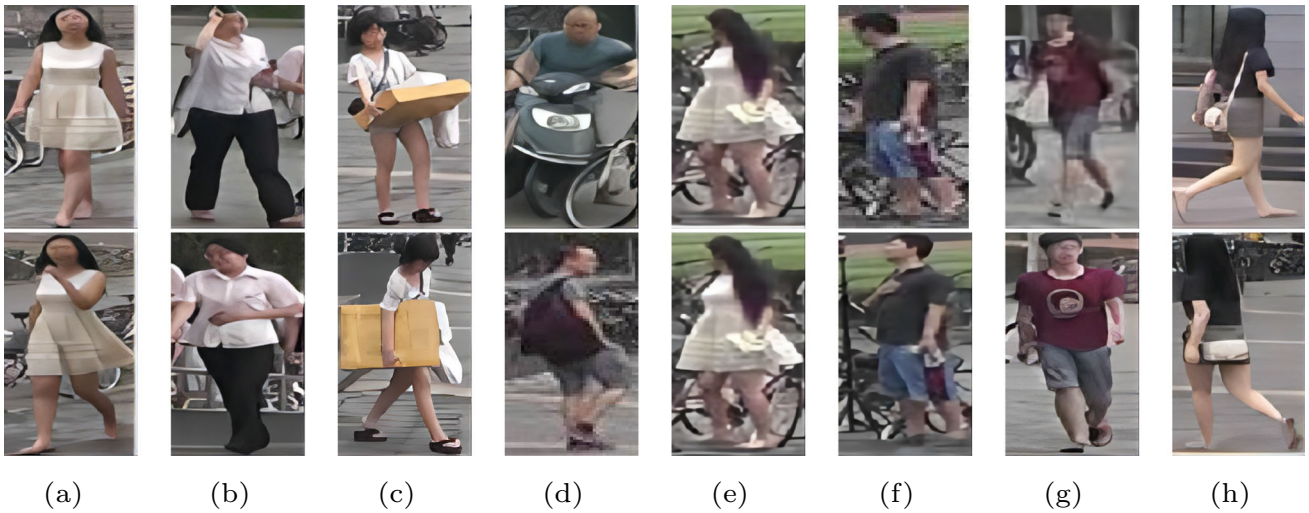

Fig. 2 The challenge of re-ID dataset samples. Each column demonstrates a sample from a unique identification. Columns **a** and **b** show changes in posture, **c** and **d** show examples of occlusion, **e** and **f** describe background clutter, and **g** and **h** show view changes in samples. All samples are from the Market1501 datasets

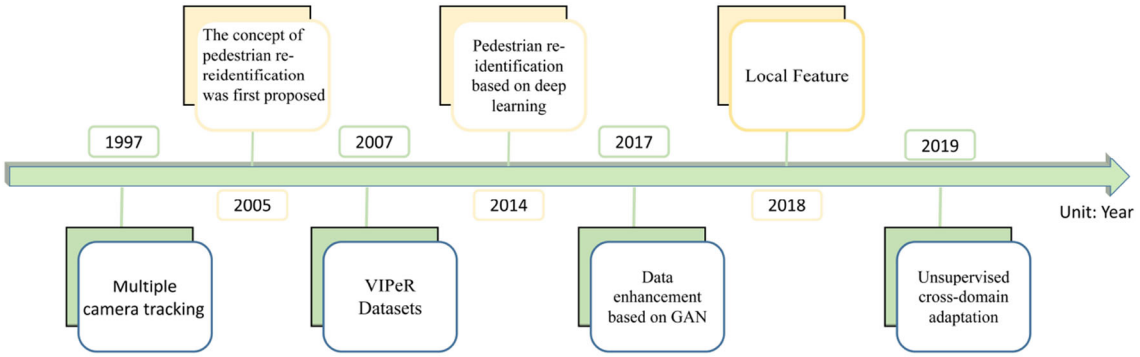

Fig. 3 Key technologies in the development of pedestrian re-ID

application of deep learning methods, and insightful discussions on future development directions. The primary work can be summarised as follows, as seen in Fig. 4:

- We present a comprehensive overview of deep learning-based pedestrian re-ID tasks, covering problem definition, primary datasets, and standard evaluation methods, providing a solid background for this paper.
- We classify pedestrian re-ID algorithms based on different deep learning approaches and application scenarios. Additionally, we conduct a detailed review of representative pedestrian re-ID systems, discussing their underlying mechanisms, advantages, limitations, and application scenarios. Furthermore, we introduce recent advancements in pedestrian re-ID approaches and evaluate the role and effectiveness of classical methods in practical applications.
- We analyze the limitations and challenges of current deep learning-based pedestrian re-ID algorithms from various perspectives and provide suitable recommendations. Finally, we outline future trends and potential development paths in the field.

The rest of the paper is organized as follows. Section "Datasets and evaluation" describes the most typical datasets and evaluation metrics. Section "State-of-art methods for pedestrian re-identification" describes numerous re-ID approaches based on various categorization criteria used in deep learning, as well as a comparison and summary of various methods. Section "Algorithm comparison and visualization results" compares traditional algorithms for pedestrian re-ID and the visualization results. Section "inlinkFuture prospects and challengess5sps1" discusses the current obstacles and potential for pedestrian re-ID. Finally, Section "Conclusion" concludes the paper.

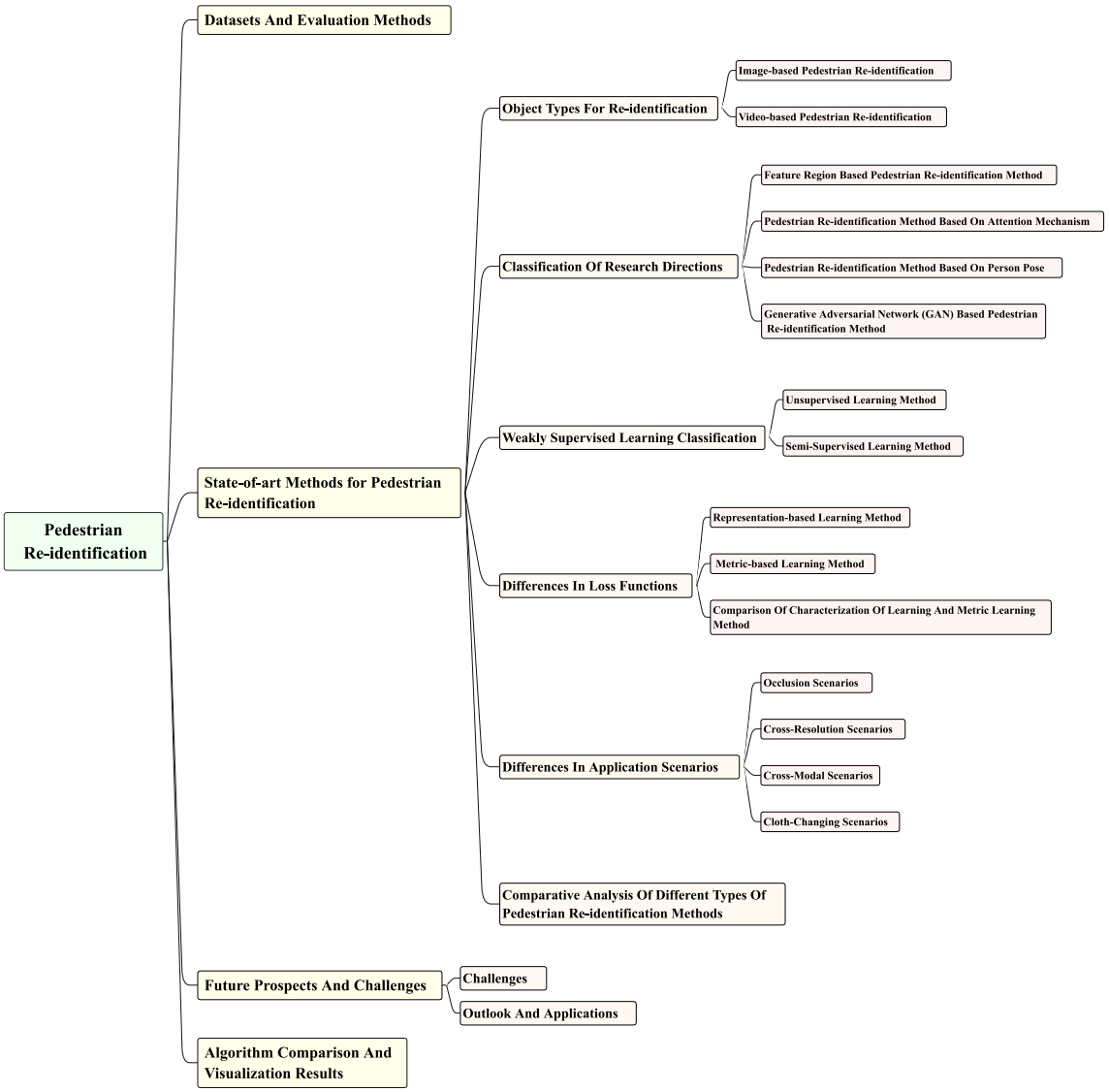

Fig. 4 Hierarchical structure of the pedestrian re-ID task

Datasets and evaluation

Datasets

This paper compiles a variety of benchmark datasets designed for evaluating the robustness and accuracy of different approaches and network models utilized in re-ID systems, specifically focusing on image and video-based re-ID.

The most frequently used single-modal datasets as well as summary cross-modal pedestrian re-ID datasets have been compiled, and algorithm performance comparisons are displayed in Table 6. Eleven commonly used single-modal datasets are given in Tables 1 and 2, including seven image datasets (VIPeR [19], iLIDS [28], PRID2011 [29], CUHK03 [22], Market-1501 [2], DukeMTMC-ReID [21], and MSMT17 [30]) and four video datasets (PRID-2011 [29], iLIDS-VID [31], MARS [32], DukeMTMC-VideoReID [33]), enumerated cross-modal pedestrian re-ID datasets, such as cross-resolution datasets, text datasets, infrared pedestrian datasets, and depth image datasets.

(1) Infrared pedestrian dataset: The SYSU-MM01 dataset [16], which includes two types of pedestrian images captured by three infrared and four visible cameras, including 491 identities of LR and RGB photos from seven cameras, providing a total of 15,792 LR images and 28,762 RGB images. RegDB dataset [34], 412 pedestrian were captured simultaneously using dual visible and infrared cameras, and the ten images captured for each pedestrian differed in lighting conditions, shooting distance, pose, and angle.

(2) Depth image dataset: The PAVIS dataset [35] contains four different sets of data. The first "collaborative" group records frontal views, extended arms, slow walking,

Table 1 Common image datasets for pedestrian re-ID

Dataset	Pedestrian	Cameras	Pictures	Methods	Evaluation
VIPeR	632	2	1264	Manual	CMC
iLIDS	119	2	476	Manual	CMC
PRID2011	934	2	24,541	Manual	CMC
CUHK03	1467	10	13,164	Manual and DPM	CMC
Market-1501	1501	6	32,217	Manual and DPM	CMC and mAP
DukeMTMC-ReID	1812	8	36,441	Manual	CMC and mAP
MSMT17	4101	15	1,26,441	Faster RCNN	CMC and mAP

Table 2 Common video datasets for pedestrian re-ID

Dataset	Pedestrian	Cameras	Videos (Bounding Box)	Methods	Evaluation
PRID-2011	200	2	400 (4000)	Manual	CMC
iLIDS-VID	300	2	600 (44,000)	Manual	CMC
MARS	1261	6	20,715 (10,00,000)	DPM and GMMCP	CMC and mAP
DukeMTMC-VideoReID	1812	8	4832 (−)	Manual	CMC and mAP

and unobstructed views of 79 pedestrian. Group 2 ("Walk 1") and group 3 ("Walk 2") data consist of the same 79 individuals. Group 4 ("Rearview") is a rear-view recording of pedestrian leaving the studio. The BIWI RGBD-ID dataset [36] collects motion video sequences of 50 different pedestrians at different locations and times.

(3) Text dataset: The CUHK-PEDES dataset [9] contains 40,206 pedestrian images of 13,003 identities. Each pedestrian image is described by two different texts, and a total of 80,412 sentences are collected, consisting of 18,93,118 words and 9408 unique words. One of the most enormous cross-modal retrieval datasets, Flickr30k [37], contains 31,783 images, each containing five textual descriptions.

(4) Cross-resolution dataset: MLR-VIPeR is constructed from the VIPeR dataset. VIPeR consists of 632 individual image pairs, all captured by two cameras. Each of these images is a high resolution of 128 × 48 pixels.

Evaluation indicators

To examine the accuracy and performance of pedestrian re-ID methods in related studies, the images of pedestrian in the database are typically separated into a training set and a test set, either arbitrarily or based on some criterion. The data detected by the first camera are used as the finding set during testing, while the personal data captured by the second camera are used as the candidate set. This section will introduce the common evaluation indicators utilized in re-ID endeavors.

(1) Rank-n accuracy

Rank-n is a widely used evaluation metric in image retrieval and classification. It is the probability of getting the correct result for the top n (highest confidence) images in the search results. An example of rank-n re-ID accuracy is shown in Fig. 5. For example, rank-1 is the ratio of the percentage of labels with the highest prediction probability to the percentage of correct labels, i.e., the percentage of correct images for the first returned image.

(2) Mean average precision

In recent years, mean average precision (mAP) [38] as an evaluation criterion can better compare the advantages and disadvantages of various approaches. The average precision is produced by charting the relationship between recall and precision (P–R curve), and the area of the curve and the coordinate axis. The calculating formula is as follows:

$$\text{Percision} = \frac{|\{\text{Image with the same ID}\} \bigcap \{\text{The query results}\}|}{|\{\text{The query results}\}|}. \tag{1}$$

To get the value of mAP, it is necessary to draw a PR curve, and then calculate the area under the PR curve to get the average accuracy AP. Therefore, the key is how to sample the PR curve. The calculation details of AP measurement were changed in 2010 [39], when there are 11 points, the maximum precision value can be selected, and AP is the average of 11 precisions. The formula is as follows:

$$\text{AP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} p_{\text{inter}}(r), \tag{2}$$

where $p_{\text{inter}}(r)$ denotes precision interpolation, and the formula is as follows:

$$\text{p}_{\text{inter}}(r) = \max p(\tilde{r}), \tag{3}$$

Fig. 5 The results are presented in Rank-n accuracy. Columns **a** represent query images, **b** represent Rank-1, **c** represent Rank-2, **d** represent Rank-3, **e** represent Rank-4, and **f** represent Rank-5. The picture with a solid line frame is successfully identified, while the picture with a dotted line frame is incorrectly identified

and the precision value is taken as the maximum of all recall $> r$. The range of $\tilde{r}$ is $\tilde{r} \geq r$. The max $p(\tilde{r})$ is the maximum measurement accuracy of recall.

The average accuracy metric reflects the model's accuracy and evaluates the ranking order given by the model results.

(3) Recall

The recall indicates how many accurate samples have been retrieved from all the retrieved sample targets, which can be expressed as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

where TP is a positive example (P) and the prediction is correct (T), while FN is a negative example (N) and the forecast is incorrect (F). In general, it is the ratio of the number of correct samples recovered to the total number of samples, which is only measured when the current category is retrieved.

(4) F-Score

The relationship between accuracy and recall is relative and must be changed based on the circumstances at hand to be effective. For instance, in certain circumstances, try to raise the accuracy rate while endeavoring to keep the recall rate the same. These two signs must be carefully considered and evaluated in many different contexts. The F-Score can

Table 3 Supervised learning performance in image datasets

Category	Algorithm	Backbone	Conditions	Venue	mAP/Rank-1(%)		
					Market-1501	DukeMTMC-ReID	CUHK-3
Global	Wang etc. [40]	CNN	Tesla K40	CVPR16	–/–	–/–	–/52-17
	SCAL [41]	ResNet50	GTX 1080Ti	ICCV19	89.30/95.80	79.60/89.00	72.30/74.80
Local	PCB and RPP [42]	ResNet50	Titan XP	ECCV18	80.90/93.30	68.10/82.90	57.50/63.70
	SONA [43]	ResNet50	GTX 1080Ti	ICCV19	88.83/95.58	78.18/89.55	79.23/81.85
	HLGAT and RR [44]	ResNet50	–	CVPR21	97.50/98.00	94.40/94.70	89.90/90.60
Auxiliary	PSE and ECN [45]	ResNet50	–	CVPR18	84-00/90.30	79.80/85.20	–/–
	AANet-152 and RR [46]	ResNet50	–	CVPR19	92.38/95.10	86.87/90.36	–/–
	VAreID and local and RR [47]	ResNet152	–	AAAI20	95.43/96.79	91.82/93.85	–/–
Backbone	BraidNet-CS and SRL [48]	SeResnext	–	CVPR18	69.48/83.70	59.49/76.44	–/88.18

help bring together precision and recall in the following ways:

$$\mathrm{F-Score} = \left(1+\beta^2\right)\frac{\mathrm{Precision\ Recall}}{\beta^2\ \mathrm{Precision}\ +\ \mathrm{Recall}}, \tag{5}$$

when $\beta = 1$, it is called $F1$-Score. At this time, the recall rate and accuracy rate have the same weight. If the accuracy rate is considered more important in some cases, then the value of β is adjusted to less than 1. If the recall rate is considered more important, the value of β can be adjusted to be greater than 1.

(5) mINP

In the multi-camera network, when the queried target pedestrian appears at multiple time points in the gallery, the ranking position that is the most difficult to match correctly determines the workload of the model. To achieve accurate tracking of the target to a greater extent, a measure with high computational efficiency, namely negative penalty (NP), can be used to measure the penalty to find the most difficult correct match

$$\mathrm{NP}_i = \frac{R_i^{\mathrm{hard}} - |G_i|}{R_i^{\mathrm{hard}}}, \tag{6}$$

where R_i^{hard} indicates the most difficult sort position to match, and $|G_i|$ represents a summary of the correct matching of query i. In general, the smaller NP represents better performance. Overall, so the average INP of the query is expressed as

$$\mathrm{mINP} = \frac{1}{n}\sum_i (1-\mathrm{NP}_i) = \frac{1}{n}\sum_i \frac{|G_i|}{R_i^{\mathrm{hard}}}. \tag{7}$$

The calculation of mINP can be successfully integrated into the CMC/mAP calculation procedure. However, the difference in mINP values will be much smaller than in smaller galleries. It can, however, accurately indicate the success of the re-ID model and can be used as a supplement to CMC and mAP indicators.

Performance of the correlation model

For visual comparison, this paper presents the performance comparison of deep learning-based pedestrian re-ID models, according to the model algorithm and the way of extracting datasets, including the performance of supervised learning on image datasets (Table 3), supervised learning on video datasets (Table 5), unsupervised learning on commonly used datasets (Table 4), and cross-modal pedestrian re-ID methods on commonly used pedestrian datasets (Table 6).

As shown in Table 3, the supervised learning pedestrian re-ID model has made significant progress on the image datasets, with Rank1 accuracy increasing from 83.7% in 2018 to 98.0% on the Market-1501 datasets improving by 14.3 percentage points. Rank1 accuracy increased from The accuracy of Rank1 on the DukeMTMC-Reid datasets increased by 18.26 percentage points from 76.44% to 94.7% in 2018.

The comparison concludes that the local feature model performs better on the datasets. However, the results achieved by different models on different datasets are also inconsistent, and researchers still need to pay further attention to the performance of the models.

As shown in Table 5, with the development of deep learning techniques, the performance of the supervised learning pedestrian re-ID model on the video datasets is improving. Specifically, on the PRID-2011 datasets, rank1 accuracy improved from 70% in 2016 to 96.2% in 2021.

On the iLIDS-VID datasets, rank1 accuracy improved from 58% to 90.4%. On the MARS datasets, accuracy improved from 44% to 91.0% in 2017.

As shown in Table 4, unsupervised pedestrian re-ID has received increasing attention, as evidenced by the number of top publications. The performance of unsupervised pedestrian re-ID models has increased significantly in recent years.

Table 4 The performance of unsupervised learning

Algorithm	Venue	Backbone	Conditions	mAP/Rank-1		
				Merket-1501	DukeMTMC-ReID	MSMT17
HHL [49]	ECCV2018	ResNet50	–	31.40/62.20	27.20/46.90	–/–
SSG [49]	ICCV2019	ResNet50	TiTAN X	58.30/80.80	53.40/73.00	–/–
SpCL [50]	NeurIPS2020	ResNet50	–	76.70/90.30	68.80/82.90	–/–
IICS [51]	CVPR2021	ResNet50	–	72.90/89.50	64.40/80.00	26.90/56.40
JVTC+ [52]	CVPR2021	ResNet50	TiTAN RTX	75.40/90.50	67.60/81.90	29.70/54.40
GLT [53]	CVPR2021	ResNet50	–	79.50/92.20	69.20/82.00	27.70/59.50

Table 5 Supervised learning performance in video datasets

Algorithm	Backbone	Conditions	Venue	mAP/Rank-1(%)		
				PRID-2011	iLIDS-VID	MARS
McLaughli etc. [54]	Siamese	GTX 980	CVPR16	–/70.00	–/58.00	–/–
ASTPN [55]	LSTM, Siamese	GTX 1080	ICCV17	–/77.00	–/62.00	–/44-00
Chen etc. [56]	ResNet50, LSTM	–	ECCV18	94.50/93.00	87.80/85.40	76.10/86.30
GRL [57]	ResNet50	Intel i4790 GTX 2080Ti	CVPR21	–/96.20	–/90.40	84.80/91.00

On the Market-1501 datasets, rank1 accuracy increased from 62.2% to 92.2% in 4 years. Performance on the DukeMTMC-Reid datasets increased from 46.9% to 82.0%. The gap between the upper bound of supervised and unsupervised learning is significantly narrowed, which proves the success of unsupervised pedestrian re-ID.

As shown in Table 6, most of the cross-modal pedestrian re-ID models in recent years are based on metric learning methods and specific feature-based models. Cross-resolution pedestrian re-ID is mainly based on unified modal methods, which are more challenging to implement for text-based pedestrian re-ID tasks. In contrast, suitable modal methods have yet to be thoroughly studied and applied.

State-of-art methods for pedestrian re-identification

Object types for re-ID

Image-based pedestrian re-ID

During the early stages, the image method was employed to build pedestrian re-ID technology. The key lies in utilizing shallow, mid-level, and deep-level visual features to explain individual appearance traits. By analyzing and comparing these features, accurate identification and matching of individuals can be achieved across different images. The integration of these visual features enhances system robustness and accuracy, enabling reliable pedestrian re-ID and tracking in various complex scenarios.

(1) Shallow visual features

Shallow features contain more detailed pixel information and have higher resolution. When the external network [69] is used to extract the shallow visual features in the image, finer-grained feature information of pedestrians in the image can be obtained from the network. At this time, the overlapping area of the perceived image field corresponding to each pixel in the feature mapping is still very small, ensuring that the network can obtain more detailed features, thus improving re-ID accuracy.

(2) Mid-level visual features (semantic attributes)

Mid-level visual features include information such as whether or not the pedestrians in the image are carrying objects, the colour of their shoes, and the length of their hair. It has less noise and more meaningful information than shallow visual features. Zhu et al. [70] used identity label alignment to re-identify personal objects and body parts at the pixel level, generating distinctive and robust representations that solved the re-ID challenge caused by pedestrian posture changes and misaligned pedestrian images. To deal with body part dislocation, Yang et al. [71] developed a semantic-guided alignment model based on image semantic attribute information to extract the more visible aspects of pedestrians in the image from occlusion noise.

(3) Deep visual features

Deep visual features are having less noise influence, a bigger receptive field, and stronger semantic information than shallow and medium visual features. Karanam et al.

Table 6 Performance comparison of cross modal pedestrian re-ID algorithms

	Datasets	Algorithm	Backbone	Conditions	Venue	Rank-1/5/10 (%)
Infrared	SYSU-MM01	Wu etc. [16]	ResNet	–	ICCV2017	24.43/–75.86
		D-HSME [58]	AlexNet	–	AAAI2019	20.68/–/62.74
		Hi-CMD [59]	ResNet50	TiTAN Xp	CVPR2020	34.94/–/77,58
		DDAG [60]	ResNet50	–	ECCV2020	61.02/–/94.06
		NFS [61]	ResNet50	2080Ti	CVPR2021	70.03/–/97.70
	RegDB	TONE [62]	–	–	AAAI2018	16.87/–/34.03
		D-HSME [58]	–	–	AAAI2019	50.85/–/73.36
		Hi-CMD [59]	–	–	CVPR2021	70.93/–/86.39
		NFS [61]	–	–	CVPR2021	80.54/–/91.96
Deep image	PAVIS	4DRAM [7]	LSTM	–	CVPR2016	43.00/–/–
		WU etc. [63]	–	–	IEEE2017	71.74/88.46/–
		Karianakis etc. [8]	LSTM	–	ECCV2018	52.40/–/–
	BIWI RGBBD-ID	4DRAM [7]	LETM	–	CVPR2016	45.30/–/–
		Karianakis etc. [8]	–	–	ECCV2018	50.00/–/–
Text	CUHK-PEDES	GNA-RNN [9]	VGG16, LSTM	–	CVPR2017	19.05/–/53.64
		Chen etc. [10]	ResNet50, LSTM	–	ECCV2018	43.58/66.93/76.26
		CMPM-CMPC [64]	ResNet152, LSTM	GTX 1080	ECCV2018	49.37/–/79.27
		A-GAN [65]	LSTM	TiTAN XP	ACMMM2019	53.14/74.03/81.95
	Flickr30k	CMPM-CMPC [64]	–	–	ECCV2018	37.30/65.70/75.50
		A-GANet [65]	–	–	ACMMM2019	39.53/69.91/80.91
Cross-resolution	MLR-VIPER	CSR-GAN [66]	ResNet50	–	IJCAI2018	37.20/62.30/71.60
		CAD-Net [67]	ResNet50	GTX 1080	ICCV2019	43.10/68.20/77.50
		MRJL [68]	RRN	–	IJCAI2021	58.70/84.10/–
	MLR-Market-1501	MRJL [68]	–	–	IJCAI2021	90.10/95.60/–

[13] encoded the attributes of different pedestrians in images sparsely and compute color histograms and textures of different stripes in the images. This method effectively addresses the challenges posed by the uniqueness and viewpoint variations in re-identification tasks. By introducing dependency aggregation module and adaptive attention mechanism, Si et al. [72] enhanced the model's understanding of spatial dependencies between images and within images, and then improved the learning ability of re-ID model. Yang et al. [73] proposed a feature mining approach that integrates pose and appearance features to enhance the discriminative capability of fused features in the re-ID model. McLaughlin et al. [54] used CNN to extract depth visual features from video. This method can capture the valuable movement and appearance information of pedestrians in the video, and improves the performance of the pedestrian re-ID system.

Image-based re-ID algorithms explore different visual features in their research. For example, Girshick et al. [69] described and classified pedestrians using locally detailed features. Zhu et al. [70] employed various semantic attribute information to describe the appearance of pedestrians. Meanwhile, McLaughlin et al. [54] extracted feature representations of pedestrian images using deep CNNs. These three algorithms differ in terms of feature selection and extraction methods. Therefore, if high-quality images and fine-grained re-ID results are required, shallow feature algorithms can be chosen. In complex scenarios, mid-level feature algorithms may have an advantage. For large-scale datasets and scenarios with high-accuracy requirements, deep feature algorithms can provide more powerful semantic representations. Additionally, incorporating methods such as feature fusion and attention mechanisms can further enhance the performance of pedestrian re-ID systems.

Video-based pedestrian re-ID

Considering the development of CNN's application in image-based re-ID, some researchers have extended it to video processing. The video-based pedestrian re-ID technology has advanced significantly in recent years, with its research concentrating mainly on constructing strong feature representations [19].

(1) Traditional methods

Traditional methods for video-based pedestrian re-ID can be categorized into manual feature-based approaches [74] and learning the method of reliable distance metrics [75].

For instance, Wang et al. [74] addressed the uncertainty and visual ambiguity in re-ID by selecting the most distinctive video clips from incomplete or noisy pedestrian image sequences. You et al. [76] proposed a pushing distance learning model that utilizes pushing constraints to match pedestrian characteristics in videos. The model also selects distinguishing features to differentiate individuals, addressing the issue of fuzzy inter-class differences in re-ID tasks.

(2) Deep learning methods

A typical video-based re-ID system in deep learning approaches consists of three components: an image-level feature extractor, a temporal modeling approach for aggregating temporal information, and a loss function [54]. For example, McLaughlin et al. [54] used a convolutional neural network (CNN) to extract pedestrian features from each frame of the video, and then integrated all of the information from each frame from the time pool to generate the overall appearance features of pedestrian targets. Hou et al. [77] used distinct attention modules to detect different body sections of pedestrians in continuous frames, obtaining the entire features of identifying the target body and significantly reducing the calculation amount of the re-ID model. Aich et al. [78] proposed a decomposition method for spatio-temporal representation, which learns separate representations for the temporal and spatial aspects, leading to improved performance of the re-ID baseline architecture. Bai et al. [79] proposed the integration and distribution module (IDM), which aims to broaden the attention region and integrate features in the feature space. The method can provide a more comprehensive feature description and help to enhance the discriminative ability of pedestrian re-ID systems. Yao et al. [80] modeled the semantic relationship between local blocks of pedestrians in video frames using complementary information and weighted sparse graphs, which significantly enhances re-ID performance. Chen et al. [81] used the spatial relation module to detect the salient regions at the video frame level, avoiding information redundancy across frames and capturing critical information, increasing re-ID robustness. Lu et al. [82] employed the time series of bone information to describe the high-order correlation between different areas of the body, improving the robust representation of re-ID based on temporal and spatial characteristics.

In conclusion, traditional methods in video-based pedestrian re-ID primarily focus on feature extraction and distance metrics. For example, You et al. [76] utilized pushing distance to learn discriminative features by matching pedestrian feature in videos. On the other hand, deep learning methods extract feature representations from videos through end-to-end learning. For instance, Yao et al. [80] modeled semantic relationships in video frames using weighted sparse graphs and complementary information, which enhances re-ID performance. As a result, deep learning methods have an advantage in modeling and exploiting temporal information and usually achieve better performance in most cases. However, traditional methods still have certain advantages in specific scenarios, particularly when dealing with smaller scale datasets or lower quality data. Therefore, the selection of an appropriate algorithm should be comprehensively considered based on the specific application scenario and requirements.

Classification of research directions

In recent years, re-ID technology based on deep learning has advanced significantly, yielding a plethora of technological techniques and directions for a wide range of application scenarios. This section focuses on several essential technologies, such as feature areas, attention mechanisms, pedestrian's postures, and building countermeasure networks, according to various study directions.

Feature region-based pedestrian re-ID method

The feature region-based re-ID method divides the input pedestrian image into horizontal stripes or several homogeneous parts. This allows for efficient observation of the different values of each partitioned feature after segmentation and accurate localization of each part by optimizing the content consistency of the segmented region. This approach has demonstrated robustness in handling partially occluded pedestrian images and pedestrian images with small-scale changes in pose in the previous studies [19, 20, 83–85].

However, the aforementioned methods have limitations as they do not adequately consider global and partial occlusion of a pedestrian, as well as pose misalignment. Consequently, they are sub-optimal when it comes to re-ID matching in arbitrarily aligned pedestrian image selections. This suggests that there may be challenges posed by large pedestrian pose variations and unconstrained automatic detection errors.

The feature region-based re-ID approaches are addressed in three scenarios based on the aforementioned concerns and challenges: horizontal stripe partitioning, local feature, and local–global feature collaboration.

(1) Horizontal stripe partitioning

In early re-ID techniques, learning local features is performed and computed by algorithms that manually produce partitioned features. For example, Gray et al. [19] utilized image segmentation to extract texture and implemented color features in re-ID tasks. This approach helps capture fine-grained texture details and color characteristics, enhancing the accuracy of re-ID. Similar partitioning methods have been used in many studies [20]. Fu et al. [86] employed the pyramid pool method to acquire feature stripes at various scales, a crucial aspect in enhancing feature robustness for re-ID tasks. This approach accurately segments pedestrian body features and mitigates the issue of misalignment. Zhu et al. [87] solved

inaccurate detection in re-ID by employing adaptive fringe and foreground thinning for robust pixel-level partial alignment.

Researchers have made significant efforts in terms of strategic methods and partitioning algorithms in the deep learning-based approach to horizontal stripe partitioning. These methods are largely summarised below.

(1) Automatic localization-based approach: Li et al. [88] used multi-scale convolution to obtain local background information of pedestrians. The method combines the global and local body part representation learning process into a unified framework, which solves the re-ID problems such as background confusion. Sun et al. [43] enhanced the uniform segmentation in the part-based convolutional baseline (PCB) network. By accurately positioning the pedestrian body parts in the image, this method improved the performance of the re-ID task.

(2) An attention-based approach: The attention mechanism to construct aligned [89–91] feature regions are achieved by suppressing background noise and enhancing the feature representation of discrimination regions. However, these methods cannot play their full role in explicitly locating semantic parts. Therefore, both Liu et al. [92] and Zhao et al. [93] implemented the model to decide the focus location by itself through an attention mechanism embedded in the network, which enriched the final feature representation of pedestrian images. In addition, efforts have been made to improve re-ID accuracy by local block matching [94]. Ning et al. [95] obtained the distinctive features of diversity and discrimination in the image through the difference of attention to improve the performance of the re-ID model.

(3) Additional semantic-based approaches: Many studies employ additional semantic methods to accurately locate body parts based on positions or postures [15, 96, 97]. Their aim is to achieve pixel-level alignment, allowing for the depiction of characteristic areas in re-ID tasks. Inspired by the twin neural network, Yi et al. [98] introduced the depth measurement learning method, which utilized a double CNN to divide the target pedestrian's image into three parts. This method enhances the accuracy of the re-ID model by evaluating the similarity between two images. Kalayeh et al. [99] proposed a pedestrian analysis model that integrates semantic analysis of individuals into the re-ID task, thereby improving the performance of re-ID.

Each of these three approaches has a specific purpose in overcoming the challenges in pedestrian re-ID. The automatic localization-based algorithms addresses the problems associated with background confusion, while the attention-based algorithms emphasizes the discriminative power of discriminative regions. In addition, the semantic-based algorithms focus on precise localization and feature representation of body parts to make re-ID more accurate.

(2) Local features

Local features are similar to horizontal stripes, which extract features from a certain area in the image and finally fuse multiple local features as the final feature. Some deep convolutional neural network (DCNN) models [98] use rigid body parts to obtain local features of pedestrians, thus improving the robustness of re-ID. From another point of view, local features can also be obtained by other methods. For example, Wang et al. [100] combined global features with local features and used local features to focus the global information in the original image. By capturing more fine-grained pedestrian features in the global information, the model can improve the accuracy of the re-ID task. Zhang et al. [101] coded the local parts of pedestrians and added a feature refinement layer, improving the re-ID model's discrimination ability. Xie et al. [102] employed a convolutional product to capture local similarity features in body and face images. This approach refined these features to calculate the final similarity, successfully incorporating facial clues into re-ID tasks. On the other hand, Xi et al. [103] introduced a powerful global feature called comprehensive global embedding. This approach enhances the differentiation of local areas within the global feature map, enabling the re-ID model to extract more fine-grained local features.

Local maximum occurrence features can obtain the horizontal occurrences of local features [20] to analyze and maximize their occurrences to represent the re-identified images. As shown in Fig. 6, local features refer to feature extraction of an image's area. Finally, multiple local features are fused to obtain the final features.

Generally speaking, these algorithms show different strategies for using local features, such as combining global and local information, refining local features, merging key clues, or using powerful global representation. Each method plays a vital role in enhancing the distinguishing ability and accuracy of the re-ID model.

(3) Local global feature collaboration

A lot of research work focuses on learning the global characteristics of re-ID [46, 104, 105]. Some recent studies study global features by designing loss functions. For example, Zheng et al. [104] enhanced the performance of the re-ID model by jointly training verification loss and classification loss during model training. This approach effectively utilized re-ID annotations to improve model performance. To obtain global image features, Hermans et al. [105] introduced a complex negative mining strategy with triple selection. This method can help to learn a more robust image representation and improve the accuracy of the re-ID task. Yang et al. [106] proposed a two-branch CNN architecture to capture both global and local features of pedestrians. By combining these features with the triple loss function, this method improves the learning efficiency of re-ID tasks.

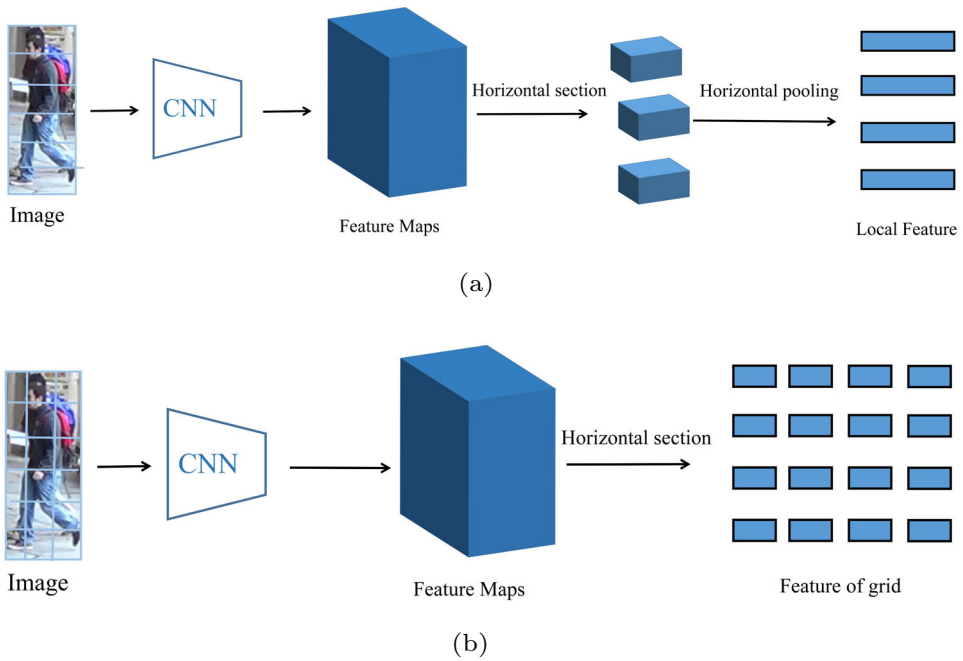

Fig. 6 Common methods of local feature division: **a** stands for horizontal pooling. The process is to divide feature maps horizontally and then pool to obtain local features of blocks; **b** Represents grid feature. The process is to take the C-dimension feature of each pixel in the feature maps of H×W×C as a grid feature. Finally, there are H×W grid feature vectors; the dimension of each vector is the number of channels C

Su et al. [107] developed a pose-driven deep convolutional model that enhances robust feature representation from local and global images using cues from pedestrian body parts. The model provides a more reliable solution for re-ID tasks in complex scenes. At the same time, Li et al. [88] improved the performance of the re-ID task by superimposing multi-scale convolution and background-aware mechanisms to extract robust pedestrian global and local features.

In summary, different algorithms in feature region-based pedestrian re-ID methods have adopted their unique strategies to enhance performance. For instance, Zheng et al. [104] and Yang et al. [106] both focused on jointly training the loss functions to improve re-ID performance. However, Zheng et al. [104] emphasized the joint training of verification and classification losses, utilizing the labeled information in re-ID. In contrast, Yang et al. [106] introduced a dual-branch CNN architecture to capture global and local features of pedestrians and train the model using a triplet loss function. This approach leverages the combination of global and local features to enhance learning efficiency.

Generally speaking, the strategies of jointly training loss functions, introducing complex negative mining strategies, extracting multi-scale features, and utilizing pedestrian body part cues have all contributed to the performance improvement in pedestrian re-ID. Selecting the appropriate algorithm based on specific application requirements and scenarios can enhance the accuracy and robustness of pedestrian re-ID.

Pedestrian re-ID method based on attention mechanism

Recent research has shown that the application of attention learning in re-ID tasks can enhance re-ID features, suppress irrelevant features, and improve the robustness of re-ID tasks. By focusing on important local features, attention mechanisms can overcome issues such as occlusion and LR [89, 108–110] in pedestrian re-ID, thereby improving accuracy and efficiency.

This section classifies attention-based re-ID methods into structured attention and multiple attention mechanisms. Also, it divides attention mechanisms into temporal and spatial attention mechanisms according to the different attention focuses.

(1) Structured attention

Some studies have found that developing well-structured patterns in the global network environment can be challenging. This is because one of the characteristics of the attention mechanism is to learn by convolution with a limited number of acceptance domains.

One approach to tackle this issue is to utilize a large-scale filter [111] in the convolution layer. This enables the filter to correspond to distinct body parts of pedestrians, thus offering valuable information on the structure of the pedestrian body for the re-ID model. Another solution involves increasing the network size to enhance the re-ID model's performance. This can be accomplished by stacking deep layers [112], thereby enabling the model to learn more intricate and discriminative features. Wang et al. [112] improved the re-ID model by adding convolution layers to the attention module, enabling it to capture more background information and enhance robust-

ness against noise labels. At the same time, Li et al. [113] enhanced the diversity and distinguishability of parts in re-ID by incorporating an attention module into the end-to-end coding and decoding structure.

(2) Multi-headed attention mechanism

To improve the comparability and stability of attention learning in the network model, multi-head attention technology is often used. For example, Ye et al. [60] used multi-headed attention to enhance the characterization of contextual relationships between visible and infrared pedestrian re-ID channels. This method improves the robustness and re-identification ability of the re-ID model for noisy images by learning different local features. Li et al. [114] proposed the harmonious attention CNN model to optimize pedestrian feature representation in images. By combining soft pixel attention and hard region attention, this model improved re-ID performance for individuals in misaligned images. He et al. [115] used a multi-headed attention mechanism to address long-distance dependence. It is able to capture more key feature information in pedestrian images and improve the accuracy and robustness of the re-ID task.

(3) Temporal attention and spatial attention mechanisms

Regarding the re-ID problem, attention mechanisms can be classified into two types based on their attention focus on features in the model: temporal attention mechanisms and spatial attention mechanisms. The binary mask is considered a type of spatial attention that can assist the model in extracting the region of interest. Song et al. [14] improved the performance and robustness of the re-ID model by focusing on pedestrian regional features and modeling the video sequences in time. In video-based re-ID problems, applying the temporal attention mechanism is more suitable. Li et al. [110] proposed a spatio-temporal attention model that effectively tackles the challenge of unaligned and occluded pedestrian regions in re-ID video sequences by fusing image features and temporal attention mechanisms. As shown in Fig. 7, Yan et al. [116] generated a multi-attention network by spatial division of features and dynamic weight distribution, which improved the distinguishing ability, robustness and representation ability of re-ID method.

Based on the attention mechanism, we can obtain the following summaries for pedestrian re-ID algorithms: (1) Structured attention-based algorithms primarily aim to capture well-structured patterns in the global network context by employing large-scale filters and increasing the network size. (2) Multi-head attention techniques, through the use of multiple attention heads, help improve attention learning and feature representation. (3) Temporal and spatial attention mechanisms address challenges related to temporal dynamics and spatial localization, respectively.

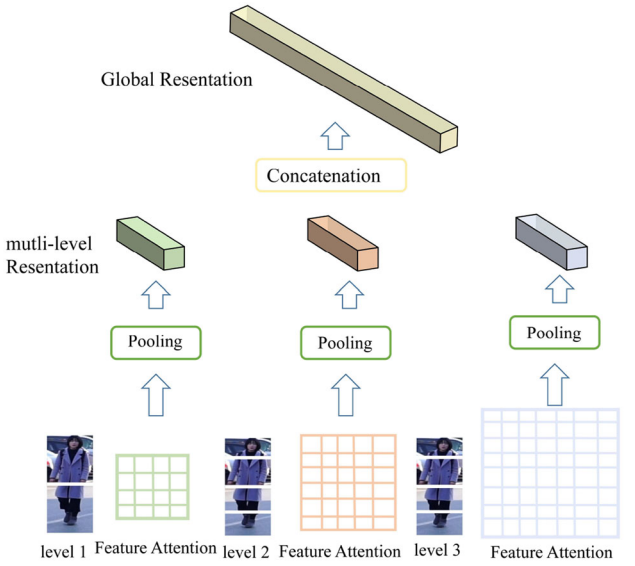

Fig. 7 Feature attention network based on re-ID for generating different levels of features in [116]

Pedestrian re-ID method based on pedestrian pose

Based on local features and attention mechanisms are two important components of re-ID technology. However, extracting features and matching pedestrian from pedestrian's images are more difficult because of various pose changes and different shooting angles in the data set.

The pose-based method effectively learns feature representations by combining the overall structure of the global image with various local parts. It helps overcome the challenge of pedestrian feature deformation caused by pose changes. In the context of re-ID research, this section focuses on the pose-based classification method, which addresses two main aspects: the misalignment problem and the pose variation problem.

(1) Misalignment problem

The solvable approaches proposed by most of the research work for problems in positional misalignment can be broadly classified into two categories: matching [94, 117] and positional attention mechanisms [70].

(1) Matching: Matching strategy and classification can be considered the critical points of matching-based approaches. Different definitions of matching components and corresponding matching strategies have been proposed to solve the location misalignment problem. These methods can be summarized as reconstruction-based matching [94, 117] and set-based matching [118].

Matching based on reconstruction involves generating and reconstructing a local feature map from the whole image with the same pedestrian identity. For example, He et al. [117] proposed a method that utilizes FCN to reconstruct a cor-

responding size local feature map, improving the similarity and accuracy in re-ID tasks.

The matching method based on set does not need the explicit alignment operation of feature space, and it regards the occluded re-ID as an integrated matching task. Jia et al. [118] calculated the similarity between pedestrian image pattern sets by introducing the Jaccard similarity coefficient as a measure, improving the re-ID task's robustness in chaotic scenes.

He et al. [117] and Jia et al. [118] addressed the misalignment problem in re-ID using different methods. The former focuses on reconstructing local feature maps, while the latter utilizes set similarity to address related issues. The choice of strategy may depend on the specific feature of the dataset and the nature of misalignment.

(2) Location attention mechanism: By learning relevant attention, the method based on the attention mechanism tackles the problem of location mismatch. The strategies for dealing with positional misalignment based on attention mechanisms can be described as clustering-based [70] and self-supervised [119] techniques, according to the important points in learning attention.

The primary way cluster-based approaches supervise attention learning is by generating pseudo-labels through clustering. Specifically, Zhu et al. [70] tackled the problem of location mismatch in re-ID by introducing cascaded clustering. It generates pedestrian site pixel-level pseudo-labels step by step and uses them to guide site attention learning. This approach effectively resolves the issue of pedestrian position mismatch in re-ID. Sun et al. [119] defined rectangular regions on the global image and extracts random patches within these regions. By assigning region labels to each pixel within the patches, it captures fine-grained information and improves re-ID accuracy.

The above two algorithms provided different strategies for handling the misalignment problem. The former tackles misaligned pedestrian locations by introducing cascaded clustering. The latter improves re-ID accuracy by capturing fine-grained information within random blocks of the global image. The choice of method may depend on the availability of annotated data and the desired level of supervision.

(2) Pose problem

Due to the restricted accuracy of the posture estimation algorithm and the effects of elements, such as occlusion, lighting changes, and complicated backgrounds, the pose estimation technique may need to be improved. Su et al. [107] proposed an enhanced end-to-end feature extraction and matching model, which can represent and match pedestrian images more accurately and improve the performance of re-ID tasks. Gu et al. [120] generated adversarial networks by feature extraction to learn pose-independent pedestrian feature representations. This approach is able to reduce the computational cost and improve the similarity matching accuracy of the same pedestrian in the re-ID model. Also, Qian et al. [121] improved the robustness and re-identification performance of the re-ID model using synthetic images to learn depth features that are not affected by changes in pedestrian pose. The network model is shown in Fig. 8.

Pose-based methods aim to create more distinguishable pedestrian pose features by considering pose variations, thereby enhancing the robustness of the re-ID model in complex conditions. For example, Su et al. [107] captured discriminative cues of pedestrian parts in images by extracting enhanced features. Gu et al. [120] utilized a feature extraction generative adversarial network to learn pose-independent representations. Both algorithms attempt to improve feature representation by incorporating pose information or learning pose-invariant representations. The choice of method may depend on the balance between accuracy and computational efficiency.

Generative adversarial network-based pedestrian re-ID method

Using the re-ID method based on deep learning mentioned in the previous part of this paper, can solve the problems of occlusion and dislocation in re-ID. However, the performance of the same method on different data sets is inconsistent. Most of these problems can be enhanced by improving GAN correlation method, thus reducing the problem of model over-fitting caused by domain gap.

One major challenge in current GANs' research is the quality of sample generation, with or without semi-supervised learning. Researchers are actively working on improving the network or model [122] to enhance the output quality of generated samples. For this reason, Zheng et al. [21] proposed a simple semi-supervised method. This approach used the original training set and enhanced the discriminative power of the re-ID model by combining the generated unlabeled data with the labeled training data.

Second, data annotation in the context of re-ID is both expensive and tedious, as it requires assigning identity labels to each pedestrian bounding box in the input images. Two factors have contributed to recent advances in this area: (1) the availability of datasets for large-scale re-ID efforts [22]. (2) Pedestrian embedding learned using CNNs [123]. Nevertheless, the number of images per identity is still limited. Therefore, it is essential to use additional data to avoid model over-fitting. To mitigate the risk of over-fitting, Zheng et al. [21] suggested employing GANs for generating unlabeled data and incorporating label smoothing regularization in the unlabeled dataset. This approach effectively reduced the burden and costs associated with data annotation in re-ID tasks.

In addition, Zhao et al. [124] added training data to the generated image samples by jointly training GAN and re-ID models, which solved the problem of insufficient re-ID data.

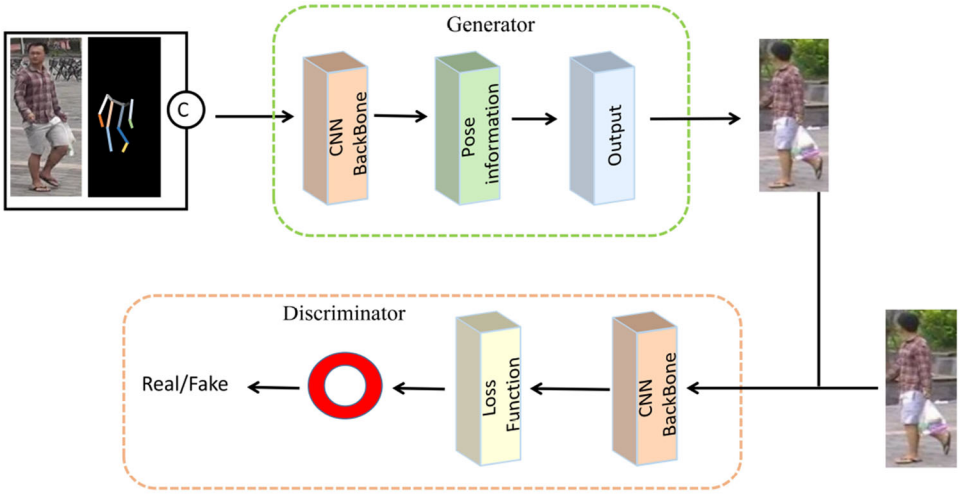

Fig. 8 Schematic of PN-GAN model [121]. Generator: given an input pedestrian image and a target pedestrian image containing pedestrian with the same ID but different poses, the generator will learn the pose information in the target pose and generate the pose information. Discriminator: designed to learn whether an input image is real or fake (i.e., binary classification task)

Zhang et al. [125] proposed a comparison learning framework that utilizes the camera center of mass as a clustering agent, which reduces the correlation between the camera and the features, thus improving the accuracy and robustness of the re-ID task. There are similarities and differences between these approaches. First, both methods offered effective solutions to key challenges, but with different focuses. The former emphasized addressing the issue of insufficient data, while the latter focused on achieving unsupervised re-ID and reducing correlation. The choice of which method to use in addressing the problem may depend on specific application requirements and data conditions.

Consequently, GAN-based algorithms in pedestrian re-ID show great promise in overcoming the challenges related to data scarcity, domain gap, and over-fitting. By generating synthetic data and incorporating advanced regularization techniques, these algorithms provide a means to effectively utilize data and reduce the reliance on extensive manual annotation, thereby contributing to improved re-ID performance.

Weakly supervised learning classification

Supervised models are able to achieve high accuracy using training data with labeled identity information for model training. In contrast, weakly supervised models use training data with weakly supervised signals for training, such as bounding boxes or image-level labels, for feature learning and matching. In this subsection, we categorize weakly supervised learning into two subcategories: unsupervised learning and semi-supervised learning.

Unsupervised learning method

While, early unsupervised re-ID mainly learns invariant components, i.e., dictionaries [126], metrics [127], etc. Deep unsupervised re-ID can be summarized into the following two categories.

The first category is unsupervised domain adaptation. It uses transfer learning to improve unsupervised re-ID [49, 53, 128], i.e., learning re-ID models from labeled source and unlabeled target domains. Zheng et al. [129] suppressed the appearance of noisy samples by clustering the confidence of the pseudo-label assigned to each sample obtained, thus mitigating the effect of noisy labels during the re-ID training process. Zhong et al. [49] obtained association information between the source and target domains by sampling training pairs using camera invariance and domain connectivity. The method improves the accuracy and generalization ability of the re-ID model in the target domain. Zheng et al. [53] proposed the group-aware label transfer (GLT) algorithm. It supports online interaction and mutual promotion of pseudo-label prediction and representation learning. This approach effectively narrows the gap between unsupervised and supervised re-ID performance. Dai et al. [130] used a hybrid mechanism based on voting, which made the source domain network maintain the distinguishability of each domain feature and improved the correlation between the source domain and the target domain of the re-ID data set. He et al. [131] introduced a pseudo-label refinement method aimed at improving the consistency of the re-ID model and eliminating the influence of noise. Zheng et al. [132] dynamically generated pseudo-labels of online samples through hierarchical clustering, which accurately reflected the true semantics of unlabeled samples and achieved better pseudo-labels and re-ID accuracy. Dai et al. [133] proposed a self-paced contrastive learning framework. At its core is the continuous and effective supervision provided by the hybrid memory model under dynamically changing categories. The model improves the performance and adaptability of re-ID through effective supervision and adaptive training. The model is shown in Fig. 9.

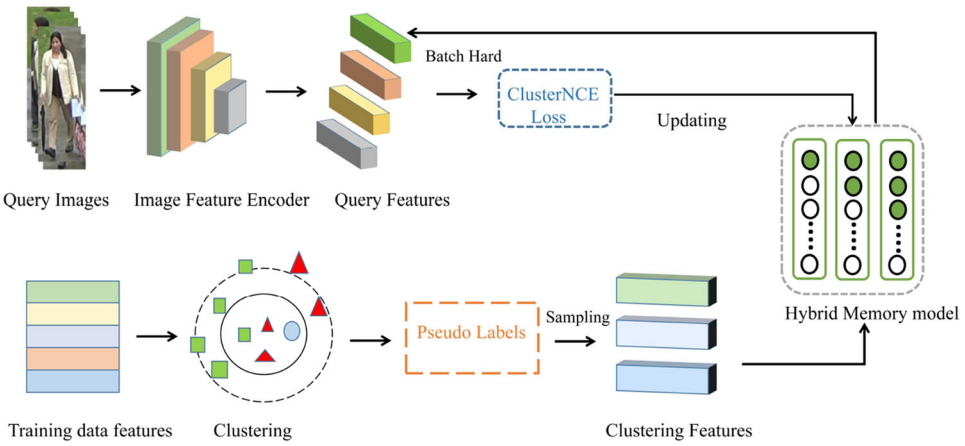

Fig. 9 The self-paced contrastive learning framework proposed in [133]

The second category is pure unsupervised learning [134–136]. Fan et al. [134] proposed a progressive unsupervised learning (PUL) method that transferred pre-trained depth representation to the invisible domain, improving the accuracy of re-ID through enhanced distinguishing features. Yang et al. [135] proposed a weighted linear coding method as an unsupervised method to learn multilevel descriptors from the original pixel data, which made the re-ID task have good robustness and uniqueness. Lin et al. [136] treated each sample as a cluster and employed a gradual grouping process to generate pseudo-labels by grouping similar samples together. This approach reduced the computational cost of the re-ID task and enhanced its accuracy. Li et al. [137] proposed an asymmetric comparative learning method guided by clustering. It improved the clustering results of unsupervised re-ID by learning discriminant features based on the clustering outcomes. Chen et al. [138] proposed to replace traditional data augmentation methods with generative adversarial networks, which generated augmented views for contrastive learning with improved performance on id-sensitive re-ID tasks. Si et al. [139] addressed the re-ID task by considering unsupervised instance-level and clustering-level feature relationships. This method generated pseudo-labels for heterogeneous images using clustering and improved feature relationships by reducing inter-modal differences with instance-level constraints. Chen et al. [140] proposed a data expansion and label assignment strategy that enhances the specificity of each semantic feature domain, leading to more reliable pseudo-labels in re-ID.

Semi-supervised learning method

Several research works on semi-supervised studies on re-ID [134, 141–143]. Yang et al. [141] explored the complementary information shared by multiple cores to effectively combine the multi-core embedding technology into the semi-supervised framework, greatly enhancing the re-ID performance. Huang et al. [142] used pseudo-regularized labels in a semi-supervised manner to enhance the correlation between the generated data classes in re-ID and the real data classes. This not only improves the robustness of the generated data, but also improves the performance of the re-ID model. Xin et al. [143] proposed a semi-supervised feature representation framework by introducing more unlabeled data for semi-supervised learning. This method can improve the robustness and generalization of pedestrian feature representation, resulting in more accurate re-ID cross-camera matching under different environmental conditions.

In summary, in the aforementioned unsupervised domain adaptation algorithms, some methods improve accuracy and alleviate the issue of noisy labels through clustering and pseudo-labeling. Other methods achieve high-accuracy re-ID results by selecting training samples based on camera invariance and domain connectivity to bridge the gap between the source and target domains. Pure unsupervised learning algorithms enhance clustering results and feature relationships through progressive grouping, clustering-guided contrastive learning, and instance-level constraints. They may also utilize generative adversarial networks to generate reliable pseudo-labels and enhance the generalization capability of re-ID models. Semi-supervised approaches further leverage limited labeled data and a large amount of unlabeled data to improve re-ID performance. Future studies can explore further integration and cross-pollination of these algorithms to achieve more accurate and scalable pedestrian re-ID systems.

Differences in loss functions

The approaches described above can solve the re-ID problem in a variety of instances. However, the algorithm's advancement is mostly apparent in the design of the loss function, which serves as a guide in the overall network optimization. According to the distinct loss functions, this section is separated into representation-based and metric-based learning approaches.

Representation-based learning method

Representation learning (RL)-based approaches are commonly used for re-ID [20, 83]. Although re-ID aims to make the network model learn the similarity between pairs of images, the representation learning-based approach does not directly consider the similarity between images when training the network model. Still, it treats the re-ID task as a classification problem or a validation problem.

(1) Classification of losses

Multi-classification problem networks generally use the softmax function as the last layer of the neural network and then calculate the cross-entropy loss or identity loss.

Cross-entropy loss: Crossover entropy loss is generally present in both binary and multi-classification problems. It captures the difference between the predicted probabilities of the network models, which in turn can measure the effectiveness and performance of different classifiers. Typically, we use y_{ij} to indicate whether the ith sample belongs to class j. y_{ij} has only two values: 0 or 1. If the output is 1, it belongs, and vice versa, if it is 0, it does not. Moreover, p_{ij} denotes the probability value that the i th sample is predicted to be the j th class, and takes the value range [0, 1]. The expression of cross-entropy loss is

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N} y_{ij}\log\left(p_{ij}\right), \quad j = 1, 2, 3, \ldots, C, \tag{8}$$

where N denotes the number of samples and C represents the total number of categories. In [84], images are fed into a classifier consisting of a fully connected (FC) layer and a softmax function in a PCB network and performed classification prediction. The PCB network is optimized by performing the sum of cross-entropy losses of ID prediction. Wu et al. [144] proposed a multicenter softmax loss to correct for head camera bias, improved the performance of the re-ID model by improving the discrimination of camera samples using a center mining strategy.

Identity loss: Identity loss treats re-ID as a classification task. Given an input image x_i labeled y_i, the predicted probability of x_i being identified as class y_i is encoded by a softmax function, denoted by $p_{i,j}$. The expression for identity loss is

$$\mathcal{L}_{id} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(y_i \mid x_i\right)\right), \tag{9}$$

where n denotes the total number of training samples in each batch, and identity loss has been widely used in existing methods [43, 92].

In [19], the network is jointly trained using identity loss and verification loss, allowing the network to learn a fused feature representation that combines both the pedestrian's identity information and visual features. This network can automatically extract features suitable for re-ID tasks and can be used to perform re-ID on new images during the testing phase. Wei et al. [85] combined an identity loss function and an online complex mining triplet loss function as a baseline learning objective. This learning objective is applied to dual-stream network for learning re-ID features and aims to improve the identifiability of the re-ID baseline.

(2) Verification loss

Verification Loss (VL) optimizes the pairwise relationship between images, i.e., input a pair of pedestrian images and let the network learn whether these two images belong to the same pedestrian, equivalent to a binary classification problem (yes or no).

We use $p(\delta_{ij} \mid f_{ij})$ to denote the probability that the input pair (x_i, x_j) is identified as δ_{ij} (0 or 1). The verification loss of cross-entropy is

$$\begin{aligned}\mathcal{L}_{\text{veri}}(i, j) = &-\delta_{ij}\log\left(p\left(\delta_{ij} \mid f_{ij}\right)\right) \\ &-\left(1-\delta_{ij}\right)\log\left(1-p\left(\delta_{ij} \mid f_{ij}\right)\right).\end{aligned} \tag{10}$$

Verification loss is frequently paired with the identity loss function to improve the performance of a model or network [89, 93].

Huang et al. [17] used classification/identification loss and verification loss to train the network, whose network schematic is shown in Fig. 10. The network input comprises several pedestrian images, including the classification and verification subnet.

The classification subnet predicts the IDs of the images and calculates the classification error loss based on the predicted IDs. The verification subnet fuses the features of two images to determine whether the two images belong to the same pedestrian, essentially equal to a binary classification network. After training with enough data, a test image is input again, and the network will automatically extract a feature used for the re-ID task.

Today, there is still a large amount of work based on representational learning methods, and representational learning has become an essential baseline in the field of re-ID. Representational learning methods are more robust, and the training process is more stable.

Metric-based learning method

Many metric learning methods have also been applied to the re-ID problem in deep learning [145]. For example, the output feature vectors from the same pedestrian in the network are closer than those from different pedestrian. These metric learning methods mainly include local fisher discriminant analysis [145] and marginal fisher analysis [104].

Different from representation learning, metric learning aims to learn the similarity between two images through neu-

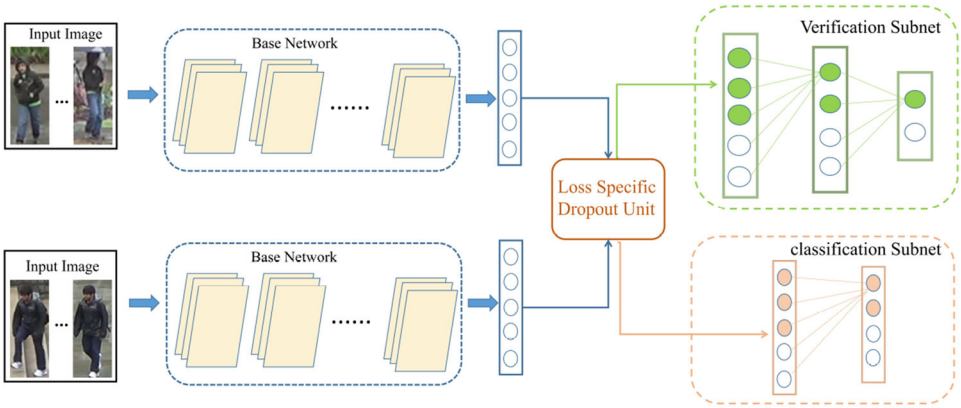

Fig. 10 Deep re-ID network framework [17]

ral networks. Define a mapping

$$f(x) : \mathbb{R}^F \rightarrow \mathbb{R}^D, \tag{11}$$

where $\mathbb{R}^F$ is the image space, $\mathbb{R}^D$ is the feature space, and $f(x)$ is the network model we want to learn.

The images are mapped from the original domain to the feature domain, after which a distance metric function is defined as follows:

$$D(x, y) : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}. \tag{12}$$

This distance metric function calculates the distance between two feature vectors.

European distance:

$$d_{I1,I2} = \| f_{I1} - f_{I2} \|_2 . \tag{13}$$

Cosine distance:

$$d_{I1,I2} = 1 - \frac{f_{I1} \cdot f_{I2}}{\| f_{I1} \|_2 \| f_{I2} \|_2}. \tag{14}$$

Because the distance function is continuous, training an end-to-end network model is possible. There are several types of learning loss functions, the most common of which are contrast loss, ternary loss, improved ternary loss, and quadruple loss.

(1) Contrastive loss

It improves the pairwise relationship of images or data. The relative two-by-two distance comparison with the expression is improved by contrastive loss

$$\mathcal{L}_c = y d_{Ia,Ib}^2 + (1 - y) \left(\alpha - d_{Ia,Ib} \right)_+^2 , \tag{15}$$

where $d_{I1,I2} \| f_{I1} - f_{I2} \|_2$ is the Euclidean distance and $(Z)_+$ denotes $\max(z, 0)$.

Several variants of the form such as the softmax-based form of the contrastive loss function called InfoNCE used in [146] to describe the similarity of the dot product metric. The contrastive loss function can also be based on other forms [147, 148], such as marginal-based loss and variants of NCELoss, which effectively improve the performance of re-ID algorithm. Zhao et al. [149] proposed a two-layer contrastive learning framework, which increased the robustness of the re-ID model by mining inter- and intra-instance similarities to reduce repulsion due to differences in the instances.

(2) Triplet loss

The triplet loss function is a loss function that is more widely used in current re-ID networks and is commonly used in face re-ID tasks. The advantage of the ternary loss is detail differentiation, i.e., when two input images are extremely similar, the ternary loss can model them based on the details of the images. The triplet loss with boundary parameters is represented as

$$\mathcal{L}_{\text{tri}} (i, j, k) = \max \left(\rho + d_{ij} - d_{ik}, 0 \right), \tag{16}$$

where $d()$ means the Euclidean distance between two samples. To solve the optimization problem of the triplet loss function, some methods [12, 105] obtain various information utilizing ternary mining, and the basic idea is to select informative triplet losses [105]. In particular, to reduce the triple loss of no information [12]. As a result, the ultimate fusing of the loss functions of each level in each step accelerates the training of the re-ID model. Liu et al. [150] improved the accuracy and robustness of the re-ID model by combining triple loss with sampling training in the metric feature space. Hermans et al. [105] moderated positive mining with weight constraints is introduced to train the robustness of CNNs for re-ID, giving the re-ID model better generalization ability.

To learn more distinguishing features, Cheng et al. [123] proposed an improved triple-state loss function that balances inter- and intra-class constraints and L2 parameter distances by adding weights. This allows the re-ID model to learn richer feature information. Zhu et al. [151] proposed combining the new supervision signals with the original softmax loss for pedestrian re-ID. As shown in Fig. 11, there are two stages

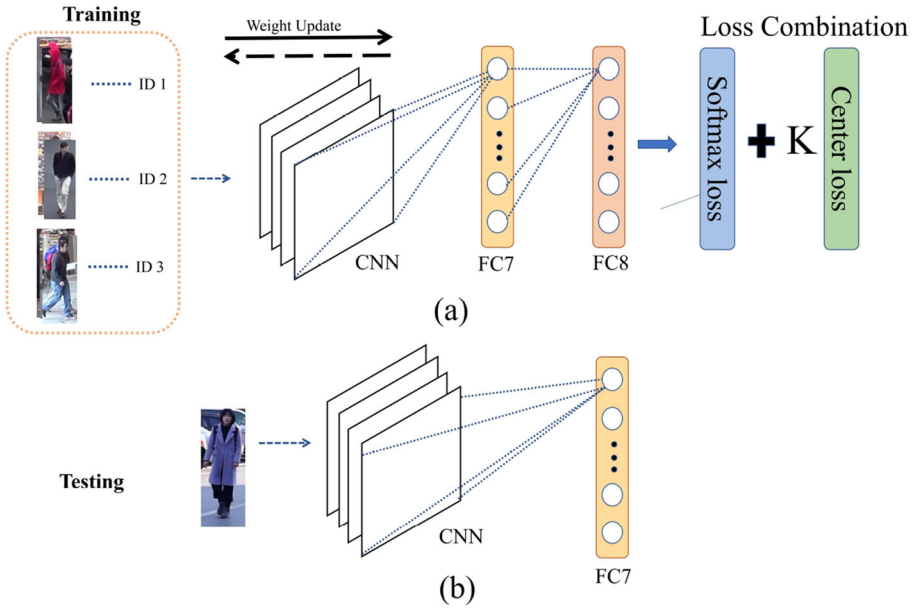

Fig. 11 CNN training model based on the combination of softmax and Center loss [151]. K adjusts the weight of the center loss, $K \in [0, 1]$. If $K = 0$, only the softmax loss is used to train the CNN model, while larger values of K emphasize the compactness of feature vectors

for [151]: (1) During the training process Fig. 11a, pedestrian images are processed through a CNN to generate feature maps. Then, these feature maps are fed into fully connected layers (FC7 and FC8) to generate feature vectors representing pedestrian identities. Finally, the hybrid loss function (combining softmax loss and center loss) is used to train the CNN model with enhanced discriminative power. During the back propagation process, the weights are updated to further enhance the feature extraction capability of FC7. This process aims to optimize the network model by adjusting the weights and improving the prediction accuracy. (2) During the testing process Fig. 11b, image feature maps are used as input to the CNN and propagated forward to FC7. The purpose of FC7 is to extract better feature representations for accurate prediction.

(3) Quadruple loss

To further enrich the triplet supervision, Chen et al. [152] proposed a quadruplet depth network, each containing one anchor sample a, one positive sample p, and two mined negative samples n_1, n_2. Where n_1 and n_2 are the IDs of two different pedestrian images. The quadruple loss is represented as

$$\mathcal{L}_q = \left(d_{a,p} - d_{a,n1} + \alpha\right) + \left(d_{a,p} - d_{n1,n2} + \beta\right)_+, \quad (17)$$

where α and β are the normal numbers set manually and usually set β less than α. The former term is called a strong push, and the latter is called a weak push. Quadruple depth network mainly produces a strong push between positive samples and negative samples, because it can make the intra-class variance smaller and the inter-class variance larger.

Comparison of characterization of learning and metric learning method

Representational learning and metric learning have their own characteristics. Ye et al. [60] combined the different roles of representation learning and metric learning in the training and testing phases to continuously optimize the network. The advantage is that the network can learn both the distance metric of the feature space and the knowledge of identity classification in the training phase, thus improving the performance of re-ID. In the testing phase, identity classification can be performed by extracting features and using the trained classifier.

Differences in application scenarios

The future application scenarios of re-ID in smart cities will encounter various challenges. The above methods proposed in this paper can effectively solve the problems encountered in multiple scenarios. According to the application of different scenes, this section is divided into occlusion scene, cross-resolution, cross-modal scene, and dressing change scene.

Occlusion scenarios

In real-world re-ID applications, pedestrians often get partially or completely occluded while passing through surveillance cameras. This occlusion hinders the capture of complete pedestrian information by the surveillance devices. Consequently, handling such noisy information has become a significant challenge in the field of re-ID. The strategies for processing noisy information are classified as noise-assisted

models [15, 153, 154] and noisy attention mechanism [119, 155, 156].

(1) Noise-assisted models

Auxiliary model-based strategies are focusing on identifying associated noise and suppressing noise creation. These methods can be further categorized into pose-based [15, 154] methods and resolution-based [6, 153] methods based on the type of supplementary model used. The pose-based approach utilizes an external pose estimation model to predict pose information, effectively separating valuable information from occlusion noise. In particular, the technique described in [153] improves the matching rate in re-ID by selectively matching unobstructed pedestrian body parts using attitude estimation data.

Similarly, ACSAP [154] used the confidence of attitude estimation to determine the visibility of the horizontal segmentation part of the image, which improves the stability of the re-ID model. Wang et al. [157] proposed a feature erasure and diffusion network to enhance re-ID model robustness by generating accurate occlusion masks and diffusing visible features. Zhang et al. [158] introduced an attitude change perception method using learned attitude transfer images and models for identity re-ID, addressing pedestrian attitude variations. Shi et al. [159] utilized advanced semantic information to alleviate occlusion issues in re-ID by mining non-occluded areas through attribute feature unwrapping.

The resolution-based methods [153] to suppress noise occlusion are achieved by employing a parsing mask estimated by an artificial parsing model. Specifically, TSA [6] utilized the external analysis output from dense pose estimation [160] to guide the learning of visible regions, thereby suppressing pedestrian-blocked areas in images and generating a visible signal. It improved the performance of the re-ID task by improving the accuracy and robustness of the model for pedestrian identity in the presence of occlusion. Lin et al. [153] used an analytic mask as a query in the self-attention mechanism to reduce the occluded noise in the image. The method improves the robustness and overall performance of the re-ID model for occlusion situations.

In the mentioned algorithms, both ACSAP [154] and TSA [6] methods utilized the results of an external pose estimation model to guide the learning of visible regions and suppress occlusion noise. On the other hand, Zhang et al. [158] focused on enhancing the model's perception of pose variations by learning pose-transformed images. These methods aim to address noise issues in re-ID and improve the robustness and accuracy of the models. However, there are some differences among these methods. ACSAP [154] emphasized the use of confidence from pose estimation to determine visibility, while TSA [6] suppressed occlusion noise through parsing masks. On the other hand, Zhang et al. [158] focused on learning pose-transformed images to improve re-ID accuracy under pose variations. The choice of which method to use may depend on specific application requirements, data conditions, and the importance placed on pose variations and occlusion noise.

(2) Noisy attention mechanism

The attention mechanism-based approach does not require additional information. According to the key points and main ideas of the attention learning process, the solutions to the noise problem based on the attention mechanism can be broadly summarized as data enhancement [119, 155, 156] and relationship-based methods [161].

(1) Data enhancement: To achieve the goal of excluding noisy occlusions, data enhancement methods [7, 155, 156] can be used to train the network and the model. This allows the network and the model to focus on the unoccluded pedestrian body parts in the images. Zhuo et al. [156] constructed an occlusion simulator, which used random blocks (patches) in the background of the source image as occlusion to solve the partial occlusion problem in the re-ID. VPM [119] extracted regional features from input images and assigns corresponding regional labels, improving the discrimination ability of the identification task. Hou et al. [162] proposed a spatio-temporal complementary network (STCnet) to recover the appearance of the occluded parts and accurately used spatio-temporal information to improve the performance of video-based re-ID. Xu et al. [163] proposed the visibility graph for computing the similarity of visible regions in two images. This method used the feature set of K nearest neighbors to recover complete features, addressing the loss of pedestrian information caused by noise interference and occlusion during feature matching.

(2) Relationship-based approach: The relationship-based method reduces occlusion interference by mining relationships between regions, refining extracted features. Specifically, OCNet [161] leveraged grouped convolutions and attention mechanisms to extract region features and utilized relational weights to refine these features, suppressing occluded or irrelevant information in the images. This approach contributed to enhancing the performance and robustness of re-ID models. Somers et al. [164] used target body part-based features and attention maps to obtain fine-grained information about pedestrians, making the masked re-ID task more efficient.

In summary, when dealing with occlusion scenarios, auxiliary noise-related algorithms can leverage additional information or models to improve re-ID performance, but they may require more complex computations and information integration. On the other hand, noise attention mechanism algorithms are more flexible and simple, as they do not require additional information, but their effectiveness may be limited in complex noise scenarios. Therefore, selecting the appropriate method based on specific application scenarios and requirements is important.

Cross-resolution scenarios

Recent research shows that some methods can alleviate the influence of pedestrian posture change, background noise, and partial occlusion on feature extraction and matching to a certain extent. However, due to the camera performance and the difference between the camera and the interested pedestrian, the captured pedestrian image usually has different solutions.

Many pioneering methods have been proposed [165] to explore and develop the common feature representation space of HR and LR images. Later, several research works were designed [166, 167], and super-resolution (SR) technology was introduced into the cross-resolution reconstruction problem.

For example, Jiao et al. [166] combined SRCNN and re-ID networks into a frame as a resolution restoration module, significantly improving the quality of LR images in re-ID. Ledig et al. [167] and Cheng et al. [123] improved the re-ID framework by incorporating an SR recovery module based on SR-GAN, optimizing the system for enhanced performance.

Some approaches employ GANs to refine the framework further. Specifically, Wang et al. [66] improved the image quality using SR-GAN in a cascaded structure. This method enhanced the performance of identity re-ID and improved the accuracy and stability of re-ID. Li et al. [67] developed a GANs' network that addressed cross-resolution reconstruction in re-ID. This approach utilized improved adversarial learning to recover lost information from LR images, improving re-ID performance. Recently, Cheng et al. [168] optimized the joint SR re-ID framework, improving compatibility between sub-networks by leveraging the knowledge of the image SR and re-ID association. This improved the image quality and the accuracy of feature extraction, further enhancing the performance of re-ID. Zhang et al. [169] improved re-ID efficiency by introducing an attention mechanism that restored the resolution of LR images, reducing the feature distribution gap between LR and HR images.

In summary, the utilization of SR techniques and GANs has shown promising results in mitigating the challenges of cross-resolution re-ID. These algorithms effectively enhance the quality of LR images and bridge the gap between LR and HR feature representations. By integrating SR-GAN modules or attention mechanisms, the generated HR images or restored LR images provide more discriminative information for accurate re-ID. Therefore, the integration of SR and GAN-based algorithms enhances the discriminative power of LR images and reduces the gap in feature distributions. However, further research is needed to address challenges posed by pose variations, occlusions, and camera differences to achieve more robust and accurate re-ID results in real-world scenarios.

Cross-modal scenarios

Unlike occlusion and multi-resolution situations, cross-modal re-ID task refers to the matching problem of different types of personal data, so cross-modal re-ID is more challenging and practical than general re-ID work. The main cross-modal scenarios are shown in Fig. 12.

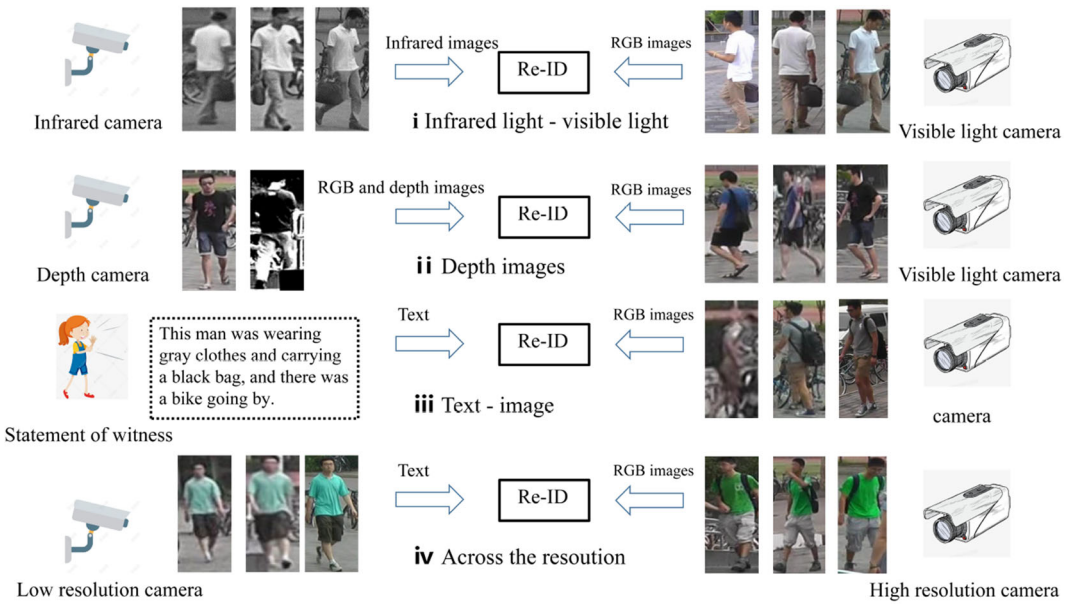

Fig. 12 Illustration of re-ID using mutli-modal and LR pedestrian data

(1) Visible-IR pedestrian re-ID

In the application of real-life scenes, an urgent problem to be solved is the re-ID problem captured across visual and infrared modes.

Early research mainly focused on two methods: representation-based learning and metric-based learning.

(1) Representation-based learning approach: This method emphasizes the development of a practical network model framework capable of capturing shared features between two distinct modal images and comparing their similarities. To solve the two problems of inter-modal and intra-modal changes across modes, Li et al. [9] solved the problem of re-ID changes under different modes by embedding mode-specific information into a shared public space using an independent CNN.

Previously, Karianakis et al. [8] minimized the cross-modal differences by exploiting the sample similarity in each modality. This method allowed better integration and understanding of information from different modalities, thus improving the overall performance and accuracy of multi-modal re-ID systems. In addition, Ye et al. [62] transformed two different modes into the same space by measuring specific modes, improving the discrimination and modal invariance of the re-ID method in complex scenes. Sun et al. [43] extracted robust pedestrian features by considering the intrinsic relationship between RGB images and IR images. This method improved the feature consistency and robustness in cross-modal cases, thus enhancing the accuracy and performance of re-ID. Gong et al. [170] tackled color transformation over-fitting in re-ID by fusing contour and color features using a local transformation attack method.

(2) Metric-based learning approach: This method aims to learn the similarity between two input images through a depth network or model. Chen et al. [10] incorporated the traditional cross-entropy loss and focused on samples that span different modalities but belong to the same property, improving the effectiveness and feature generalization of the cross-modal re-ID task. Liu et al. [65] introduced a feature learning framework for cross-modal re-ID that effectively handles cross-modal and intra-modal changes. This approach improved the accuracy of the cross-modal re-ID model by incorporating identity loss into the framework.

(3) Inter-modal based approach: Compared with the representation-based and metric-based learning methods, the GAN network can solve the problem of the significant difference between modes. For example, to effectively reduce the differences between modes, Wang et al. [171] solved the remarkable modal differences between RGB and IR images by distinguishing mode-specific features from mode-invariant features. Similarly, Dai et al. [172] improved the performance and robustness measurement of re-ID methods by combining features from different modalities and leveraging the advantages of each modality, enabling it to better adapt and identify targets in different environments and conditions. Hao et al. [173] minimized inter-modal differences and maximize cross-modal instance similarity for improved cross-modal re-ID discriminability. Liu et al. [174] proposed a memory-enhanced one-way metric learning method that enhances cross-modal correlation and mitigates modal differences, addressing the cross-modal re-ID problem. Alehdaghi et al. [175] bridged RGB and IR modalities using intermediate virtual domains to provide additional information for deep re-ID models.

In summary, future research can explore the combination and integration of these algorithms to improve the accuracy and robustness of visible-IR re-ID. For example, the combination of representation-based learning and metric-based learning methods can be explored to capture shared features and learn inter-modal similarities effectively. Additionally, addressing challenges such as pose variations, occlusions, and modal differences caused by different cameras remains crucial to achieve accurate and robust visible-IR re-ID results in real-world scenarios.

(2) Pedestrian re-ID based on depth image

In the task of cross-modal re-ID, pedestrian skeleton information and body shape characteristics are captured by depth images, which provides the possibility for re-ID in poor light, changing clothes and blurred scenes. Haque et al. [7] enhanced the robustness of re-ID models to variations in viewpoint, lighting, and pose by leveraging the motion features and shape contours of individuals in images, along with the combination of CNN and recurrent neural networks (RNN). Some methods [176, 177] are dedicated to combining RGB images and depth image information. For example, Yang et al. [176] combined the appearance features of RGB-based images with the estimated depth features, which can more effectively reduce the number of features caused by complex background noise in the images. Karianakis et al. [8] effectively solved the problems of small scale and poor training efficiency of deep re-ID data set using the shallow shared parameters of re-ID model between two different modes. Wu et al. [178] employed an end-to-end RGB-D identification module to reduce the difficulty of re-ID in different modes, effectively narrowing the gap between two heterogeneous images.

These algorithms all pay attention to the pedestrian skeleton information and body shape characteristics in the depth image in cross-modal re-ID to solve the problems of insufficient illumination, changing clothes and blurring the scene. However, there are some similarities and differences between these methods. For example, Haque et al. [7] combined CNN and RNN to deal with different visual angles, illumination, and posture changes, while Yang et al. [176] reduced the influence of complex background noise by combining RGB appearance features and depth features. Karianakis et al. [8] dealt with the re-ID problem in different modes by shar-

ing parameters, while Wu et al. [178] used the end-to-end RGBD re-ID module to narrow the gap between heterogeneous images. The choice of these methods may depend on the specific application requirements, data conditions, and the degree of attention to different issues.

(3) Text-image pedestrian re-ID

The matching problem between RGB images and text descriptions can be addressed by re-ID text within the images. By utilizing the Gated Neural Attention model (GNA-RNN) of an RNN, shared feature learning between pedestrian pictures and text descriptions can be achieved [9]. The principle of this approach is to enable end-to-end training for text-to-pedestrian image retrieval. It achieves this by leveraging the feedback of varying weights based on the correlation between text and image, along with a similarity retrieval target method. To enhance the semantic correlation between language and local visual features, Chen et al. [10] improved re-ID efficiency through global and local image-language associations, learning global visual features in images. Shao et al. [179] improved re-ID performance by employing a multi-modal shared dictionary approach. This approach reconstructed visual and text features and extracted distinct and semantically consistent features for both modes, enhancing re-ID accuracy.

In future research, it would be beneficial to further explore the fusion and integration of multi-modal information. For instance, jointly modeling depth images, RGB images, and textual information to capture comprehensive pedestrian features and semantic information. Additionally, addressing the challenges arising from modal differences, such as lighting variations, pose variations, and occlusions, is crucial to improve matching and cross-modal re-ID performance. Furthermore, investigating more efficient and robust model training and optimization methods to handle large-scale datasets and challenges in real-world scenarios would be valuable.

Cloth-changing scenes

In the above scenario, most assumptions are based on the fact that pedestrian will keep their dress [43, 120] in the short term. However, if we want to re-ID a pedestrian for a long time, the problem of changing clothes is inevitable. Because of the critical role of dress re-ID in intelligent monitoring systems [176, 180], it has attracted more and more attention in recent years.

Extracting clothing-independent features is the key to solving the core problem of dressing-based re-ID. Therefore, Barbosa et al. [35] separate pedestrians' appearance and structure information from RGB images and regard the structure information as a feature unrelated to clothing, bringing a new research direction for the re-ID method. Fan et al. [180] used radio frequency signals to extract more persistent pedestrian features. The algorithm captured unique features associated with pedestrian identity and was capable of coping with long-term surveillance or re-ID needs across time periods. Jin et al. [181] integrated gait re-ID technology as an auxiliary task in combination with the image re-ID stream. This approach facilitated the learning of representations that are independent of wearing, enhancing the overall performance of image re-ID. Gu et al. [182] proposed an adversarial loss based on clothes to help change clothes re-ID by obtaining features and information from the original image unrelated to clothes. Hong et al. [183] supplemented the knowledge of clothes-independent shapes and improved re-ID performance. Lu et al. [184] obtained the movement information of pedestrians by fusing gait and appearance features, generated robust and discriminating features for better identity re-ID. Wu et al. [185] used a clothing-independent spatial attention module to eliminate the interference of clothing appearance, by obtaining information features from the body resolution module, effectively reducing the computational cost in the re-ID task. Zhang et al. [186] used a validation network to calculate the similarity scores between images. This approach improved the efficiency of the clothes-changing re-ID task.

However, multi-modal methods need additional models or equipment to extract multi-modal information. Compared with the above technology, gait information has stronger consistency and reliability. To keep more time and space data, Chao et al. [187] regarded gait as a set of independent frames and used CNN to learn identity information from it, which was robust to pedestrian re-ID from different perspectives. Fan et al. [188] proposed a time-based partial gait re-ID framework, which makes use of the fine-grained local information of pedestrians, to improve the re-identification ability of re-ID. Yu et al. [189] enhanced the accuracy and robustness of re-ID by combining identity loss and triplet loss to comprehensively capture the biological features of pedestrians.

Many research directions have turned to image-based clothing re-ID methods based on the above problems. To enhance the feature learning process, Jia et al. [190] through the input image's positive and negative data enhancement, the feature learning is seamlessly enhanced without additional information, and the robustness to pedestrian clothing changes is improved. Cai et al. [191] proposed a multi-scale mask-guided attention network that can better capture the details and crucial features of pedestrians, enabling accurate re-identification of pedestrians under different environments and pose variations. Yu et al. [192] provided a semi-supervised clothing-invariant feature learning framework to generate images of clothing changes. This method implemented through discriminative embedding learning of clothing-simulating generative response networks, reducing the reliance of the re-ID model on pedestrian clothing.

In summary, some algorithms attempt to extract clothing-independent features from RGB images by separating appearance and structural information and using the structure information as features unrelated to clothing for re-ID. On the other hand, some algorithms leverage other modalities or techniques, such as radio frequency signals and gait re-ID to extract more persistent and clothing-independent pedestrian features. Additionally, there are algorithms that enhance the robustness to clothing changes by enhancing the feature learning process or utilizing global and local attention mechanisms.

Comparative analysis of different types of pedestrian re-ID methods

To extract features, the global feature approach involves feeding a pair of images into a convolution network or model. This method is practical, straightforward, and efficient. Still, it is easily impacted by noisy samples and background noise. Therefore, the global feature technique is now almost exclusively employed with other methods and is practically only utilized with others.

The local feature approach aims to extract critical information about the pedestrian in the image and fine-grained information about the image. However, it is incapable of involving the global semantic details of the image comprehensively. Currently, researchers process the image with horizontal chunking and horizontal partitioning, then align the pedestrian's various parts locally, and combine them with global features.

The image feature-based method employs diverse semantic information from the image, as well as visual features from the bottom, middle, and deep levels, to efficiently distinguish distinct pedestrians without being impacted by changes in illumination and perspective.

Video feature-based approaches, on the other hand, can extract rich pedestrian spatial dimension information, optical flow information, and time information from video sequences or through the use of models such as CNNs. Nevertheless, it takes time and consumes a large amount of hardware resources during the training stage.

The performance of supervised re-ID methods is approaching saturation. Researchers' attention has shifted to weakly supervised re-ID approaches, particularly unsupervised re-ID methods based on unclassified data, which lessen reliance on labeled data and make them more relevant and practical. Unfortunately, the influence of cheap background noise samples and the low performance of related algorithms led to the model's poor performance.

In general, cross-resolution re-ID methods involve introducing images with varying resolutions into the network for matching. More specifically, low-resolution images are converted into high-resolution images using either supervised or adversarial learning techniques. This method results in the introduction of additional noise information. Nonetheless, it improves the quality of the appearance data that the images contain.

The cross-modal re-ID approach is a popular topic in current research, primarily addressing the issue of cross-modal pedestrian feature matching. However, the model's performance needs to be improved, because features with mode discrimination properties are difficult to extract and are often influenced by unknown factors such as noisy samples.

The attention mechanism is commonly used in re-ID tasks, and its primary goal is to strengthen distinguishing traits while suppressing irrelevant ones. In re-ID tasks, RNN models are typically integrated with attention mechanisms, and attention mechanisms are employed to process specific regions with high resolution. The aim is to process certain image areas and retrieve critical information for associated regions while ignoring irrelevant information.

The GAN network can handle image complementation problems, which provides an image complementation solution to the partially occluded re-ID challenge. GAN can also help with sample generation, and its adversarial training discrimination model eliminates the difficulties of loss function design, making it useful in unsupervised and semi-supervised learning domains. The GAN network, on the other hand, is difficult to train and prone to model collapse and gradient disappearance.

In Table 7, we present nine classical forms of pedestrian re-ID methods, and characterise their relevant mechanisms, advantages, limits, and applicable scenarios, based on the preceding analysis description.

Algorithm comparison and visualization results

Algorithm comparison

To provide a more visual overview of the re-ID algorithm models discussed in Section "State-of-art methods for pedestrian re-identification" , in Table 8, we show their usage on various datasets, including Market-1501 [2], MSMT17 [30], CUHK03 [22], PRID [29], MARS [32], and DukeMTMC-reId [21]. Table 8 also includes performance results measured by metrics, such as mAP and Rank-1. The comparison and analysis of the typical algorithms in Sect. 3 are as follows:

(1) As a first remark: both PPLR [193], and PLCC [194] are unsupervised learning-based re-ID algorithms that generate pseudo-labels utilizing clustering, but the methods applied in improving pseudo-labels are different. First, when calculating the similarity between pseudo-labels or features, PPLR uses a k-nearest neighbor search method to generate a similarity ranking between global and local features by con-

Table 7 Comparison of common methods in pedestrian re-ID tasks

Category	Mechanism	Advantages	Limitations	Scenes
Global	Feature extraction of pedestrian image global information by combining attention and other methods	Simple, fast	Be easily influenced	Scenarios requiring high real-time performance
Local	Image segmentation, key point positioning and foreground segmentation are used to extract features from a certain region of the image, and finally several local features are fused	Diversity	Complex in structure	Occlusion complex scene
Image-based methods	The semantic information of the image and the low-level, middle and deep visual features are used	Strong robustness	Difficult to train	Visible light and dark scenes
Video-based methods	RNN, attention mechanism and CNN are used to extract feature vectors frame by frame, and then generate video-level feature representations through time aggregation	Wealth of pictorialand temporal information	Complex model huge datasets	Real-time video sequence monitoring
Unsupervised methods	The unlabeled data is falsified by clustering method, and then the model is trained by supervised methods	Independence	Easily influenced	Unlabeled data
Across resolution	Using super-resolution techniques and adversarial learning techniques to recover lost details from low-resolution images	Rich information	Extra noise	Samples of different resolution
Cross-modal-infrared mode	Use GAN technology to unify modes or design models to reduce the differences betweenmodes and pay attention to the information of each mode	Night condition	Image details	Night and poor light conditions
Mechanism attention	Usually, the cyclic neural network is combined with the attention mechanism to act on the specific part of the picture, the Encoder–Decoder framework	Highly targeted	Quantity of data	Background noisy scene
GAN	The interaction of a generator and a discriminator	Application range	Difficult to train	Generate the pedestrian re-ID task sample

Table 8 Comparison of classical algorithms in pedestrian re-ID task

Algorithm	Year	Key idea	Loss function	mAP/Rank-1(%)				
				Market-1501	MSMT17	CUHK03	MARS	PRID
PPLR [193]	2022	Reduce label noise by adopting a complementary relationship between global and local features	Cross entropy, triplet loss	81.5/92.80	31.4/61-10	–/–	–/–	–/–
RLCC [194]	2021	Use time propagation and integration of pseudo-labels to improve the pseudo-label	Contrastive loss	77.7/90.80	27.9/56.60	–/–	–/–	–/–
MG- CAM [14]	2018	Masked guided contrastive attention model (MGCAM) is designed to learn features from the body and background areas respectively	Triplet loss	74.33/83.79	–/–	50.21/50.14	–/–	–/–
DLPAR [93]	2017	Use the attention module to deal with body parts dislocation	Triplet loss	62-10/82.3	–/–	37.83/40.93	–/–	–/–
MAR [195]	2019	Deep model of soft multi-label learning for unsupervised re-ID	Multi-label learning loss	40.00/67.70	48.00/67.10	–/–	–/–	–/–
SP- GAN [196]	2018	Transform the marked image from the source domain to the target domain in an unsupervised way	Adversarial loss	27.10/51.50	22.3/41-10	–/–	–/–	–/–
MGH [197]	2020	By modeling spatio-temporal dependence based on multi-granularity	Cross-entropy, triplet loss	–/–	–/–	–/–	85.80/90.00	–/94.80
Adaptive graph [198]	2020	Realize the background information interaction between relevant regional features	Cross-entropy, triplet loss	–/–	–/–	–/–	81.90/91.50	–/94.60
LSP [70]	2020	Locate pedestrian body parts and personal items only at the pixel level of a pedestrian with an ID label re-ID	Cross-entropy, triplet loss	88.60/95.30	–/–	74.10/76.50	–/–	–/–
P^2-Net [96]	2020	Self-attention mechanism is applied to capture soft potential partial masks	Cross-entropy, triplet loss	85.60/95.20	–/–	78.30/73.60	–/–	–/–

sidering the complementary relationship between global and local features. PLCC evaluates the similarity between $t-1$ generation and t-generation pseudo-labels by a clustering consensus approach.

Furthermore, the PLRR uses a cross-protocol score-based system to remove noise from pseudo-labels. PLCC uses temporal propagation and contained pseudo-labels to refine the pseudo-labels but ignores the essential fine-grained information in the images. Because the PLRR method utilizes the fine-grained critical details in the pictures, PPLR outperforms PLCC on the datasets Market-1501 and MSMT17.

(2) As a second remark: Both MGCAM [14], and DLPAR [93] are local representation-based pedestrian re-ID algorithms. MGCAM is the first to use a binary mask approach to make features more robust and to augment the CNN, while DLPAR learns features by re-ranking on the ResNet-50 model, MGCAM's mAP and Rank- on dataset CUHK03 1 accuracy scores on dataset CUHK03 are at least 10% better than DLPAR. On dataset Market-150 MGCAM learns features from body and background regions separately using RGB-M as input and a mask-guided contrastive attention mechanism, while DLPAR only partitions body parts, so MGCAM's mAP and Rank-1 accuracy scores were improved by at least 3% over DLPAR.

(3) As a third remark: SPGAM [196] migrates the source domain data to the target domain in an unsupervised manner. Still, the training and test sets have overlapping categories, ignoring the discriminative label information mining in the unlabeled target domain.

MAR [195] proposed a soft multi-label learning method to guide the unpaired label learning discriminatory information in the graph library and mines the unlabeled target data in the discriminative information. Hence, MAR performs much better than SPGAM on Market-1501 and MSMT17 datasets.

(4) As a fourth remark: Both AdaptiveGraph [198] and MGH [197] algorithms are video-based pedestrian re-ID methods, differing in that AdaptiveGraph requires additional pedestrian pose information to construct the adaptive graph and ignores long-term temporal dependencies by only considering correlations between adjacent frames in the video.

The main improvement of the MGH framework is the hypergraph learning mechanism, which captures the dependency of spatio-temporal cues and learns more critical information by adding an attention module and mutual information loss. Therefore, MGH outperforms AdaptiveGraph on the datasets PRID and MARS.

(5) As a fifth remark: Both LSP [70], and P^2-Net [96] are improved algorithms for fine-grained information of pedestrians in images. LSP generates pseudo-labels of pedestrian parts by clustering, while P^2-Net captures potential local features by extracting binary pedestrian part masks and self-attentive mechanisms.

Because the error-prone semantics extracted by the pre-trained model on the Market-1501 and CUHK03 datasets can significantly degrade the performance of the different semantics-based approaches, LSP performs well on this dataset, demonstrating that the learned semantic part outperforms the external semantic part in terms of robustness.

Visualization results

We selected five classical unsupervised re-ID algorithms in Sect. 3, and trained them on the unlabeled Market-1501 data set. The training data results of these algorithms are shown in Fig. 13. The experimental result shows that after 30 epochs, the training accuracy tends to be flat.

In general, the longer the training period, the higher the re-ID effect and accuracy. However, this does not imply that lengthening the training period will increase model performance. It is also affected by elements such as learning rate, model network, parameters, and other variables.

In addition, we shown the visualization results of the above algorithm or model or method on the Market-1501 data set from Figs. 14, 15, 16, 17, and 18.

Future prospects and challenges

Although re-ID work has flourished as a result of deep learning, there are still numerous hurdles to overcome and various research goals that must be addressed in the future.

Challenges

(1) Semi-supervised, unsupervised modeling studies

The wide application of unsupervised and semi-supervised learning reduces the dependence on dataset labeling and facilitates the expansion of datasets. The future research direction of semi-supervised learning should focus on extracting discrimination pedestrian features using less annotated and more unannotated datasets. GAN technology can be combined with unsupervised learning to reduce the cost of labeling. When some unlabeled data are used, the effective feature mapping space of target can be found by unsupervised learning. The transformation problem of different scenes can be overcome by joint transfer learning, which makes the re-ID problem more efficient.

(2) A cross-modal pedestrian re-identification study

The variation between different modalities not only complicates the acquisition of images in datasets, but also presents a significant challenge for the application of re-ID tasks in real-world scenarios, which must be able to automatically process data and images with varying resolutions, modalities, environments, and domains. As a result, it is more

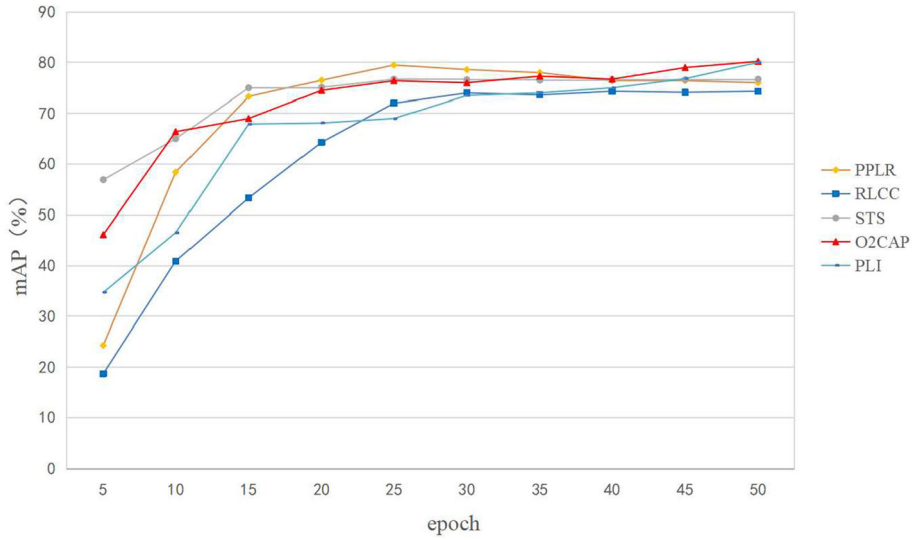

Fig. 13 Performance of different models on the Market-1501 dataset

Fig. 14 Visualization results of PPLR [193]

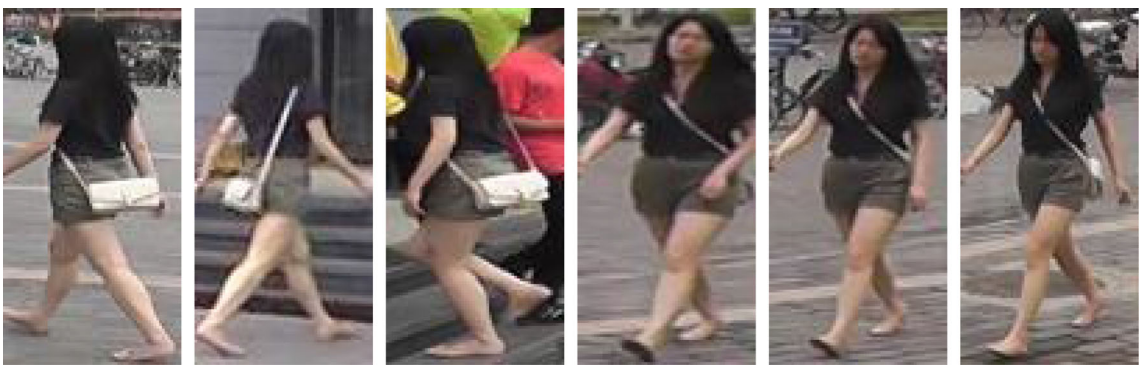

Fig. 15 Visualization results of PLCC [194]

vital to create a network and model that can withstand and adapt to various cross-domain circumstances.

(3) A study on changing clothes for pedestrian re-identification

At present, the critical points of the re-ID task to identify pedestrians are mainly focused on features, such as the face, body part information, posture, and clothes. However, in practical scenarios, it is likely to include the same pedestrian wearing different clothes, and current research on this

Fig. 16 Visualization results of STS [199]

Fig. 17 Visualization results of O2CAP [200]

Fig. 18 Visualization results of PLI [201]

situation still relies heavily on the face, gait, and spatial polar coordinate transformations as well as physical appearance to solve this problem, which may be unstable in practical applications.

Further research on re-ID techniques for individuals who alter their clothing may address this issue through additional discriminatory indicators (e.g., 3D models, etc.).

(4) Characteristic expression improvement

Extracting more critical feature expressions is the key to solving the character-heavy re-ID task. In addition, additional semantic information such as chronological order also facilitates more detailed feature representation, and models can be trained using better quality and larger datasets, which can help the model extract more valuable feature representations.

(5) Pedestrian detection combined with pedestrian re-identification

Existing re-ID tasks build datasets by cutting and labelling video segments from surveillance recordings to acquire pedestrian images, which requires extensive preparation. A distinct re-ID model is insufficient for the application's demands in actual and complex settings. Integrating pedestrian detection and re-ID has significant scientific value as well as practical implications for actual application scenarios.

Outlook and applications

In the future application of smart cities, various fields, and scenarios can employ pedestrian re-ID technology to address problems and increase the convenience of pedestrian's lives.

(1) In the scenes of crowded areas such as airports, metro, and train stations, where the flow of pedestrian is large, there are many external factors such as pose, object occlusion, and resolution difference between the images of pedestrians taken by cameras. These factors make it difficult for the extracted pedestrian features to be discriminating and robust. On the other hand, re-ID accuracy can be increased by designing a convolution neural network that incorporates crucial personal information, so that re-ID can be implemented. This will allow us to use re-ID technology to efficiently find lost children and older pedestrian, prevent mishaps involving public safety, and create a safe environment.
(2) In the realm of cross-modal scene re-ID techniques, the reliability of color as a factor determining appearance is questionable. Additionally, the extraction of supplementary information, such as body shape, can impede the accuracy of pedestrian re-ID. As a result, it becomes crucial to devise intelligent model constraints or propose innovative neural network architectures that can effectively capture and utilize body shape information, thus enhancing the cross-modal re-ID rate. This aspect also holds significance in the domain of public security, enabling the rapid screening of suspicious individuals and facilitating precise actions against criminal activities through the application of re-ID technology.
(3) In many scenes, it is possible to realize statistics of pedestrian flow by strengthening the combination of re-ID technology and pedestrian detection technology. Furthermore, it is also possible to learn the restoration of pedestrian flow trajectory, and personnel comparison for the whole scene, which facilitates real-time management and deployment of various terminal resources and helps relevant departments to deploy pedestrian resources better, improve office efficiency and optimize service experience.

With the development of intelligent cities and smart places, the convenience of pedestrian's lives increases. The widespread use of re-ID technology in numerous industries can also contribute to improving city development.

Conclusion

This paper provides an in-depth survey of the most recent research on deep learning-based approaches for pedestrian re-ID techniques. It also discusses current methodology and tools, such as standard datasets, assessment metrics, advanced pedestrian re-ID techniques, application scenario analysis, and comparison of various types of pedestrian re-ID methods. In addition, we point out potential research directions, such as integrating weakly supervised learning with neural network strategies, combining pedestrian detection with re-ID, and including additional crucial cues with discriminative features such as 3D models. The implementation of re-ID technology in real-world environments, such as smart cities, has the potential to facilitate not just the improvement of technology but also of pedestrian's quality of life. To conclude, this paper aims to assist researchers in gaining a better understanding of deep learning, re-ID technology, and its applications in some practical scenarios, hence promoting its progress and development.

Acknowledgements This research was funded by the Natural Science Foundation of Shandong Province (ZR2020QF108, ZR2022QF037, ZR2020MF148, ZR2020QF046), and the National Natural Science Foundation of China (62103350, 62072391, 62066013, 62273290).

Data availability Data available on request from the authors.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. arXiv preprint arXiv:1610.02984

2. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
3. Martinel N, Luca Foresti G, Micheloni C (2019) Aggregating deep pyramidal representations for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops
4. Gu X, Chang H, Ma B, Shan S (2022) Motion feature aggregation for video-based person re-identification. IEEE Trans Image Process 31:3908–3919
5. Dai J, Zhang P, Wang D, Lu H, Wang H (2018) Video person re-identification by temporal residual learning. IEEE Trans Image Process 28(3):1366–1377
6. Ye M, Liang C, Wang Z, Leng Q, Chen J, Liu J (2015) Specific person retrieval via incomplete text description. In: Proceedings of the 5th ACM on international conference on multimedia retrieval, pp 547–550
7. Haque A, Alahi A, Fei-Fei L (2016) Recurrent attention models for depth-based person identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1229–1238
8. Karianakis N, Liu Z, Chen Y, Soatto S (2018) Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 715–733
9. Li S, Xiao T, Li H, Zhou B, Yue D, Wang X (2017) Person search with natural language description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1970–1979
10. Chen D, Li H, Liu X, Shen Y, Shao J, Yuan Z, Wang X (2018) Improving deep visual representation for person re-identification by global and local image-language association. In: Proceedings of the European conference on computer vision (ECCV), pp 54–70
11. Ye M, Wang Z, Lan X, Yuen PC (2018) Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI, vol 1, p 2
12. Wang Y, Wang L, You Y, Zou X, Chen V, Li S, Huang G, Hariharan B, Weinberger KQ (2018) Resource aware person re-identification across multiple resolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8042–8051
13. Karanam S, Li Y, Radke RJ (2015) Person re-identification with discriminatively trained viewpoint invariant dictionaries. In: Proceedings of the IEEE international conference on computer vision, pp 4516–4524
14. Song C, Huang Y, Ouyang W, Wang L (2018) Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1179–1188
15. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 542–551
16. Wu A, Zheng W-S, Yu H-X, Gong S, Lai J (2017) Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 5380–5389
17. Huang Y, Zha Z-J, Fu X, Zhang W (2019) Illumination-invariant person re-identification. In: Proceedings of the 27th ACM international conference on multimedia, pp 365–373
18. Qian X, Wang W, Zhang L, Zhu F, Fu Y, Xiang T, Jiang Y-G, Xue X (2020) Long-term cloth-changing person re-identification. In: Proceedings of the Asian conference on computer vision
19. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision. Springer, pp 262–275
20. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2197–2206
21. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE international conference on computer vision, pp 3754–3762
22. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 152–159
23. Wu D, Zheng S-J, Zhang X-P, Yuan C-A, Cheng F, Zhao Y, Lin Y-J, Zhao Z-Q, Jiang Y-L, Huang D-S (2019) Deep learning-based methods for person re-identification: a comprehensive review. Neurocomputing 337:354–371
24. Almasawa MO, Elrefaei LA, Moria K (2019) A survey on deep learning-based person re-identification systems. IEEE Access 7:175228–175247
25. Ming Z, Zhu M, Wang X, Zhu J, Cheng J, Gao C, Yang Y, Wei X (2022) Deep learning-based person re-identification methods: a survey and outlook of recent works. Image Vis Comput 119:104394
26. Gupta A, Pawade P, Balakrishnan R (2022) Deep residual network and transfer learning-based person re-identification. Intell Syst Appl 16:200137
27. Wu D, Huang H, Zhao Q, Zhang S, Qi J, Hu J (2022) Overview of deep learning based pedestrian attribute recognition and re-identification. Heliyon 8(12):e12086
28. Zheng W-S, Gong S, Xiang T (2009) Associating groups of people. BMVC 2:1–11
29. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on image analysis. Springer, pp 91–102
30. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
31. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. In: European conference on computer vision. Springer, pp 688–703
32. Zheng L, Bie Z, Sun Y, Wang J, Su C, Wang S, Tian Q (2016) Mars: A video benchmark for large-scale person re-identification. In: European conference on computer vision. Springer, pp 868–884
33. Wu Y, Lin Y, Dong X, Yan Y, Ouyang W, Yang Y (2018) Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5177–5186
34. Nguyen DT, Hong HG, Kim KW, Park KR (2017) Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors 17(3):605
35. Barbosa IB, Cristani M, Bue AD, Bazzani L, Murino V (2012) Re-identification with rgb-d sensors. In: European conference on computer vision. Springer, pp 433–442
36. Munaro M, Basso A, Fossati A, Van Gool L, Menegatti E (2014) 3d reconstruction of freely moving persons for re-identification with a depth sensor. In: 2014 IEEE international conference on robotics and automation (ICRA). IEEE, pp 4512–4519
37. Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist 2:67–78
38. Chen Y-C, Zheng W-S, Lai J-H, Yuen PC (2016) An asymmetric distance model for cross-view feature mapping in person reiden-

tification. IEEE Trans Circuits Syst Video Technol 27(8):1661–1675
39. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
40. Wang F, Zuo W, Lin L, Zhang D, Zhang L (2016) Joint learning of single-image and cross-image representations for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1288–1296
41. Chen G, Lin C, Ren L, Lu J, Zhou J (2019) Self-critical attention learning for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9637–9646
42. Xia BN, Gong Y, Zhang Y, Poellabauer C (2019) Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3760–3769
43. 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: Proceedings of the European conference on computer vision (ECCV), pp 480–496
44. Zhang Z, Zhang H, Liu S (2021) Person re-identification using heterogeneous local graph attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12136–12145
45. Sarfraz MS, Schumann A, Eberle A, Stiefelhagen R (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 420–429
46. Tay C-P, Roy S, Yap K-H (2019) Aanet: Attribute attention network for person re-identifications. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7134–7143
47. Zhu Z, Jiang X, Zheng F, Guo X, Huang F, Sun X, Zheng W (2020) Aware loss with angular regularization for person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 13114–13121
48. Wang Y, Chen Z, Wu F, Wang G (2018) Person re-identification with cascaded pairwise convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1470–1478
49. Zhong Z, Zheng L, Li S, Yang Y (2018) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European conference on computer vision (ECCV), pp 172–188
50. Ge Y, Zhu F, Chen D, Zhao R et al (2020) Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. Adv Neural Inf Process Syst 33:11309–11321
51. Xuan S, Zhang S (2021) Intra-inter camera similarity for unsupervised person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11926–11935
52. Chen H, Wang Y, Lagadec B, Dantcheva A, Bremond F (2021) Joint generative and contrastive learning for unsupervised person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2004–2013
53. Zheng K, Liu W, He L, Mei T, Luo J, Zha Z-J (2021) Group-aware label transfer for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5310–5319
54. McLaughlin N, Del Rincon JM, Miller P (2016) Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1325–1334
55. Xu S, Cheng Y, Gu K, Yang Y, Chang S, Zhou P (2017) Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 4733–4742
56. Chen D, Li H, Xiao T, Yi S, Wang X (2018) Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1169–1178
57. Liu X, Zhang P, Yu C, Lu H, Yang X (2021) Watching you: global-guided reciprocal learning for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13334–13343
58. Hao Y, Wang N, Li J, Gao X (2019) Hsme: hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8385–8392
59. Choi S, Lee S, Kim Y, Kim T, Kim C (2020) Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10257–10266
60. Ye M, Shen J, J Crandall D, Shao L, Luo J (2020) Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: European conference on computer vision. Springer, pp 229–247
61. Chen Y, Wan L, Li Z, Jing Q, Sun Z (2021) Neural feature search for rgb-infrared person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 587–597
62. Ye M, Lan X, Li J, Yuen P (2018) Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
63. Wu A, Zheng W-S, Lai J-H (2017) Robust depth-based person re-identification. IEEE Trans Image Process 26(6):2588–2603
64. Zhang Y, Lu H (2018) Deep cross-modal projection learning for image-text matching. In: Proceedings of the European conference on computer vision (ECCV), pp 686–701
65. Liu J, Zha Z-J, Hong R, Wang M, Zhang Y (2019) Deep adversarial graph attention convolution network for text-based person search. In: Proceedings of the 27th ACM international conference on multimedia, pp 665–673
66. Wang Z, Ye M, Yang F, Bai X, 0001 SS (2018) Cascaded sr-gan for scale-adaptive low resolution person re-identification. In: IJCAI, vol 1, p 4
67. Li Y-J, Chen Y-C, Lin Y-Y, Du X, Wang Y-CF (2019) Recover and identify: a generative dual model for cross-resolution person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8090–8099
68. Zhang G, Chen Y, Lin W, Chandran A, Jing X (2021) Low resolution information also matters: learning multi-resolution representations for person re-identification. arXiv preprint arXiv:2105.12684
69. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
70. Zhu K, Guo H, Liu Z, Tang M, Wang J (2020) Identity-guided human semantic parsing for person re-identification. In: European conference on computer vision. Springer, pp 346–363
71. Yang Q, Wang P, Fang Z, Lu Q (2020) Focus on the visible regions: semantic-guided alignment model for occluded person re-identification. Sensors 20(16):4431
72. Si T, He F, Wu H, Duan Y (2022) Spatial-driven features based on image dependencies for person re-identification. Pattern Recogn 124:108462
73. Yang J, Zhang C, Li Z, Tang Y, Wang Z (2023) Discriminative feature mining with relation regularization for person re-identification. Inf Process Manage 60(3):103295
74. Wang T, Gong S, Zhu X, Wang S (2016) Person re-identification by discriminative selection in video ranking. IEEE Trans Pattern Anal Mach Intell 38(12):2501–2514

75. Zhu X, Jing X-Y, You X, Zhang X, Zhang T (2018) Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. IEEE Trans Image Process 27(11):5683–5695
76. You J, Wu A, Li X, Zheng W-S (2016) Top-push video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1345–1353
77. Hou R, Chang H, Ma B, Huang R, Shan S (2021) Bicnet-tks: learning efficient spatial-temporal representation for video person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2014–2023
78. Aich A, Zheng M, Karanam S, Chen T, Roy-Chowdhury AK, Wu Z (2021) Spatio-temporal representation factorization for video-based person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 152–162
79. Bai S, Ma B, Chang H, Huang R, Chen X (2022) Salient-to-broad transition for video person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7339–7348
80. Yao Y, Jiang X, Fujita H, Fang Z (2022) A sparse graph wavelet convolution neural network for video-based person re-identification. Pattern Recogn 129:108708
81. Chen C, Ye M, Qi M, Wu J, Liu Y, Jiang J (2022) Saliency and granularity: discovering temporal coherence for video-based person re-identification. IEEE Trans Circuits Syst Video Technol 32(9):6100–6112
82. Lu J, Wan H, Li P, Zhao X, Ma N, Gao Y (2023) Exploring high-order spatio-temporal correlations from skeleton for person re-identification. IEEE Trans Image Process. https://doi.org/10.1109/TIP.2023.3236144
83. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
84. Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B (2016) Deepercut: a deeper, stronger, and faster multi-person pose estimation model. In: European conference on computer vision. Springer, pp 34–50
85. Wei S-E, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4724–4732
86. Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang T (2019) Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8295–8302
87. Zhu K, Guo H, Liu S, Wang J, Tang M (2022) Learning semantics-consistent stripes with self-refinement for person re-identification. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2022.3151487
88. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 384–393
89. Chen B, Deng W, Hu J (2019) Mixed high-order attention network for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 371–381
90. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
91. Si J, Zhang H, Li C-G, Kuen J, Kong X, Kot AC, Wang G (2018) Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5363–5372
92. 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: Proceedings of the IEEE international conference on computer vision, pp 350–359
93. Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3219–3228
94. Zheng W-S, Li X, Xiang T, Liao S, Lai J, Gong S (2015) Partial person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 4678–4686
95. Ning X, Gong K, Li W, Zhang L (2021) Jwsaa: joint weak saliency and attention aware for person re-identification. Neurocomputing 453:801–811
96. Guo J, Yuan Y, Huang L, Zhang C, Yao J-G, Han K (2019) Beyond human parts: Dual part-aligned representations for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3642–3651
97. Liu J, Ni B, Yan Y, Zhou P, Cheng S, Hu J (2018) Pose transferrable person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4099–4108
98. Yi D, Lei Z, Liao S, Li SZ (2014) Deep metric learning for person re-identification. In: 2014 22nd International conference on pattern recognition. IEEE, pp 34–39
99. Kalayeh MM, Basaran E, Gökmen M, Kamasak ME, Shah M (2018) Human semantic parsing for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1062–1071
100. 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
101. Zhang M, Xiao Y, Xiong F, Li S, Cao Z, Fang Z, Zhou JT (2022) Person re-identification with hierarchical discriminative spatial aggregation. IEEE Trans Inf Forensics Secur 17:516–530
102. Xie Q, Lu Z, Zhou W, Li H (2022) Improving person re-identification with multi-cue similarity embedding and propagation. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2022.3207949
103. Xi J, Huang J, Zheng S, Zhou Q, Schiele B, Hua X-S, Sun Q (2023) Learning comprehensive global features in person re-identification: ensuring discriminativeness of more local regions. Pattern Recogn 134:109068
104. Zheng Z, Zheng L, Yang Y (2017) A discriminatively learned cnn embedding for person reidentification. ACM Trans Multimedia Comput Commun Appl (TOMM) 14(1):1–20
105. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737
106. Yang B, Shan Y, Peng R, Li J, Chen S, Li L (2022) A feature extraction method for person re-identification based on a two-branch cnn. Multimedia Tools Appl 81(27):39169–39184
107. Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3960–3969
108. Fang P, Zhou J, Roy SK, Petersson L, Harandi M (2019) Bilinear attention networks for person retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8030–8039
109. Fu Y, Wang X, Wei Y, Huang T (2019) Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8287–8294
110. Li S, Bak S, Carr P, Wang X (2018) Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 369–378

111. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
112. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3156–3164
113. Li Y, He J, Zhang T, Liu X, Zhang Y, Wu F (2021) Diverse part discovery: occluded person re-identification with part-aware transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2898–2907
114. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
115. He S, Luo H, Wang P, Wang F, Li H, Jiang W (2021) Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 15013–15022
116. Yan Y, Ni B, Liu J, Yang X (2019) Multi-level attention model for person re-identification. Pattern Recogn Lett 127:156–164
117. He L, Liang J, Li H, Sun Z (2018) Deep spatial feature reconstruction for partial person re-identification: alignment-free approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7073–7082
118. Jia M, Cheng X, Zhai Y, Lu S, Ma S, Tian Y, Zhang J (2021) Matching on sets: conquer occluded person re-identification without alignment. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35, pp 1673–1681
119. Sun Y, Xu Q, Li Y, Zhang C, Li Y, Wang S, Sun J (2019) Perceive where to focus: learning visibility-aware part-level features for partial person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 393–402
120. Gu X, Chang H, Ma B, Zhang H, Chen X (2020) Appearance-preserving 3d convolution for video-based person re-identification. In: European conference on computer vision. Springer, pp 228–243
121. Qian X, Fu Y, Xiang T, Wang W, Qiu J, Wu Y, Jiang Y-G, Xue X (2018) Pose-normalized image generation for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 650–667
122. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434
123. Cheng D, Gong Y, Zhou S, Wang J, Zheng N (2016) Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1335–1344
124. Zhao Z, Song R, Zhang Q, Duan P, Zhang Y (2022) Jot-gan: a framework for jointly training gan and person re-identification model. ACM Trans Multimedia Comput Commun Appl (TOMM) 18(1s):1–18
125. Zhang G, Zhang H, Lin W, Chandran AK, Jing X (2023) Camera contrast learning for unsupervised person re-identification. IEEE Trans Circuits Syst Video Technol 33(8):4096–4107
126. Elyor K, Xiang T, Fu Z, Gong S (2016) Person re-identification by unsupervised l1 graph learning. In: Proceedings of the 14th European conference on computer vision (ECCV), Amsterdam, The Netherlands, pp 8–16
127. Liu Z, Wang D, Lu H (2017) Stepwise metric promotion for unsupervised video person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 2429–2438
128. Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2019) Invariance matters: exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 598–607
129. Zheng K, Lan C, Zeng W, Zhang Z, Zha Z-J (2021) Exploiting sample uncertainty for domain adaptive person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 3538–3546
130. Dai Y, Li X, Liu J, Tong Z, Duan L-Y (2021) Generalizable person re-identification with relevance-aware mixture of experts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16145–16154
131. He T, Shen L, Guo Y, Ding G, Guo Z (2022) Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 879–887
132. Zheng Y, Tang S, Teng G, Ge Y, Liu K, Qin J, Qi D, Chen D (2021) Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In: Proceedings of the IEEE/cvf international conference on computer vision, pp 8371–8381
133. Dai Z, Wang G, Yuan W, Liu X, Zhu S, Tan P (2021) Cluster contrast for unsupervised person re-identification. arXiv preprint arXiv:2103.11568
134. Fan H, Zheng L, Yan C, Yang Y (2018) Unsupervised person re-identification: clustering and fine-tuning. ACM Trans Multimedia Comput Commun Appl (TOMM) 14(4):1–18
135. Yang Y, Wen L, Lyu S, Li S (2017) Unsupervised learning of multi-level descriptors for person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 31
136. Lin Y, Dong X, Zheng L, Yan Y, Yang Y (2019) A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8738–8745
137. Li M, Li C-G, Guo J (2022) Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. IEEE Trans Image Process 31:3606–3617
138. Chen H, Wang Y, Lagadec B, Dantcheva A, Bremond F (2022) Learning invariance from generated variance for unsupervised person re-identification. IEEE Trans Pattern Anal Mach Intell 45(6):7494–7508
139. Si T, He F, Li P, Song Y, Fan L (2023) Diversity feature constraint based on heterogeneous data for unsupervised person re-identification. Inf Process Manage 60(3):103304
140. Chen F, Wang N, Tang J, Yan P, Yu J (2023) Unsupervised person re-identification via multi-domain joint learning. Pattern Recogn 138:109369
141. Yang X, Wang M, Hong R, Tian Q, Rui Y (2017) Enhancing person re-identification in a self-trained subspace. ACM Trans Multimedia Comput Commun Appl (TOMM) 13(3):1–23
142. Huang Y, Xu J, Wu Q, Zheng Z, Zhang Z, Zhang J (2018) Multi-pseudo regularized label for generated data in person re-identification. IEEE Trans Image Process 28(3):1391–1403
143. Xin X, Wang J, Xie R, Zhou S, Huang W, Zheng N (2019) Semi-supervised person re-identification using multi-view clustering. Pattern Recogn 88:285–297
144. Wu J, Yang Y, Lei Z, Yang Y, Chen S, Li SZ (2023) Camera-aware representation learning for person re-identification. Neurocomputing 518:155–164
145. Paisitkriangkrai S, Shen C, Van Den Hengel A (2015) Learning to rank in person re-identification with metric ensembles. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1846–1855
146. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
147. Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y (2018) Learning deep representations by

mutual information estimation and maximization. arXiv preprint arXiv:1808.06670
148. Wu Z, Xiong Y, Yu SX, Lin D (2018) Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3733–3742
149. Zhao Y, Shu Q, Shi X (2023) Dual-level contrastive learning for unsupervised person re-identification. Image Vis Comput 129:104607
150. Liu D, Wu L, Hong R, Ge Z, Shen J, Boussaid F, Bennamoun M (2023) Generative metric learning for adversarially robust open-world person re-identification. ACM Trans Multimedia Comput Commun Appl 19(1):1–19
151. Zhu F, Kong X, Wu Q, Fu H, Li M (2018) A loss combination based deep model for person re-identification. Multimedia Tools Appl 77(3):3049–3069
152. Chen W, Chen X, Zhang J, Huang K (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 403–412
153. 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). IEEE, pp 2299–2303
154. He Y, Yang H, Chen L (2021) Adversarial cross-scale alignment pursuit for seriously misaligned person re-identification. In: 2021 IEEE international conference on image processing (ICIP). IEEE, pp 2373–2377
155. Chen P, Liu W, Dai P, Liu J, Ye Q, Xu M, Chen Q, Ji R (2021) Occlude them all: occlusion-aware attention network for occluded person re-id. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11833–11842
156. Zhuo J, Chen Z, Lai J, Wang G (2018) Occluded person re-identification. In: 2018 IEEE international conference on multimedia and expo (ICME). IEEE, pp 1–6
157. Wang Z, Zhu F, Tang S, Zhao R, He L, Song J (2022) Feature erasing and diffusion network for occluded person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4754–4763
158. Zhang L, Jiang N, Diao Q, Zhou Z, Wu W (2022) Person re-identification with pose variation aware data augmentation. Neural Comput Appl 34(14):11817–11830
159. Shi Y, Ling H, Wu L, Zhang B, Li P (2022) Attribute disentanglement and registration for occluded person re-identification. Neurocomputing 470:226–235
160. Güler RA, Neverova N, Kokkinos I (2018) Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7297–7306
161. Kim M, Cho M, Lee H, Cho S, Lee S (2022) Occluded person re-identification via relational adaptive feature correction learning. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2719–2723
162. Hou R, Ma B, Chang H, Gu X, Shan S, Chen X (2019) Vrstc: occlusion-free video person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7183–7192
163. Xu B, He L, Liang J, Sun Z (2022) Learning feature recovery transformer for occluded person re-identification. IEEE Trans Image Process 31:4651–4662
164. Somers V, De Vleeschouwer C, Alahi A (2023) Body part-based representation learning for occluded person re-identification. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1613–1623
165. Jing X-Y, Zhu X, Wu F, You X, Liu Q, Yue D, Hu R, Xu B (2015) Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp 695–704
166. Jiao J, Zheng W-S, Wu A, Zhu X, Gong S (2018) Deep low-resolution person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
167. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
168. Cheng Z, Dong Q, Gong S, Zhu X (2020) Inter-task association critic for cross-resolution person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2605–2615
169. Zhang G, Ge Y, Dong Z, Wang H, Zheng Y, Chen S (2021) Deep high-resolution representation learning for cross-resolution person re-identification. IEEE Trans Image Process 30:8913–8925
170. Gong Y, Huang L, Chen L (2022) Person re-identification method based on color attack and joint defence. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4313–4322
171. Wang G-A, Zhang T, Yang Y, Cheng J, Chang J, Liang X, Hou Z-G (2020) Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12144–12151
172. Dai P, Ji R, Wang H, Wu Q, Huang Y (2018) Cross-modality person re-identification with generative adversarial training. In: IJCAI, vol 1, p 6
173. Hao X, Zhao S, Ye M, Shen J (2021) Cross-modality person re-identification via modality confusion and center aggregation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 16403–16412
174. Liu J, Sun Y, Zhu F, Pei H, Yang Y, Li W (2022) Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19366–19375
175. Alehdaghi M, Josi A, Cruz RM, Granger E (2023) Visible-infrared person re-identification using privileged intermediate information. In: Computer vision–ECCV 2022 workshops, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V. Springer, pp 720–737
176. Yang Q, Wu A, Zheng W-S (2019) Person re-identification by contour sketch under moderate clothing change. IEEE Trans Pattern Anal Mach Intell 43(6):2029–2046
177. Zhang Z, Tran L, Yin X, Atoum Y, Liu X, Wan J, Wang N (2019) Gait recognition via disentangled representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4710–4719
178. Wu J, Jiang J, Qi M, Chen C, Zhang J (2022) An end-to-end heterogeneous restraint network for rgb-d cross-modal person re-identification. ACM Trans Multimedia Comput Commun Appl (TOMM) 18(4):1–22
179. Shao Z, Zhang X, Fang M, Lin Z, Wang J, Ding C (2022) Learning granularity-unified representations for text-to-image person re-identification. In: Proceedings of the 30th acm international conference on multimedia, pp 5566–5574
180. Fan L, Li T, Fang R, Hristov R, Yuan Y, Katabi D (2020) Learning longterm representations for person re-identification using radio signals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10699–10709
181. Jin X, He T, Zheng K, Yin Z, Shen X, Huang Z, Feng R, Huang J, Chen Z, Hua X-S (2022) Cloth-changing person re-identification from a single image with gait prediction and regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14278–14287

182. Gu X, Chang H, Ma B, Bai S, Shan S, Chen X (2022) Clothes-changing person re-identification with rgb modality only. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1060–1069
183. Hong P, Wu T, Wu A, Han X, Zheng W-S (2021) Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10513–10522
184. Lu X, Li X, Sheng W, Ge SS (2022) Long-term person re-identification based on appearance and gait feature fusion under covariate changes. Processes 10(4):770
185. Wu J, Liu H, Shi W, Tang H, Guo J (2022) Identity-sensitive knowledge propagation for cloth-changing person re-identification. In: 2022 IEEE international conference on image processing (ICIP). IEEE, pp 1016–1020
186. Zhang R, Fang Y, Song H, Wan F, Fu Y, Kato H, Wu Y (2023) Specialized re-ranking: a novel retrieval-verification framework for cloth changing person re-identification. Pattern Recogn 134:109070
187. Chao H, He Y, Zhang J, Feng J (2019) Gaitset: regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8126–8133
188. Fan C, Peng Y, Cao C, Liu X, Hou S, Chi J, Huang Y, Li Q, He Z (2020) Gaitpart: temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14225–14233
189. Yu S, Li S, Chen D, Zhao R, Yan J, Qiao Y (2020) Cocas: a large-scale clothes changing person dataset for re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3400–3409
190. Jia X, Zhong X, Ye M, Liu W, Huang W (2022) Complementary data augmentation for cloth-changing person re-identification. IEEE Trans Image Process 31:4227–4239
191. Cai H, Wang Z, Cheng J (2019) Multi-scale body-part mask guided attention for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops
192. 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 Multimedia 24:4482-4492
193. Cho Y, Kim WJ, Hong S, Yoon S-E (2022) Part-based pseudo label refinement for unsupervised person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7308–7318
194. Zhang X, Ge Y, Qiao Y, Li H (2021) Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3436–3445
195. Yu H-X, Zheng W-S, Wu A, Guo X, Gong S, Lai J-H (2019) Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2148–2157
196. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 994–1003
197. Yan Y, Qin J, Chen J, Liu L, Zhu F, Tai Y, Shao L (2020) Learning multi-granular hypergraphs for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2899–2908
198. Wu Y, Bourahla OEF, Li X, Wu F, Tian Q, Zhou X (2020) Adaptive graph representation learning for video person re-identification. IEEE Trans Image Process 29:8821–8830
199. Liu T, Lin Y, Du B (2022) Unsupervised person re-identification with stochastic training strategy. IEEE Trans Image Process 31:4240–4250
200. Wang M, Li J, Lai B, Gong X, Hua X-S (2022) Offline-online associated camera-aware proxies for unsupervised person re-identification. arXiv preprint arXiv:2201.05820
201. Zhang X, Li D, Wang Z, Wang J, Ding E, Shi JQ, Zhang Z, Wang J (2022) Implicit sample extension for unsupervised person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7369–7378