



Deep learning-based person re-identification methods: A survey and outlook of recent works

Zhangqiang Ming, Min Zhu ^{*}, Xiangkun Wang, Jiamin Zhu, Junlong Cheng, Chengrui Gao, Yong Yang, Xiaoyong Wei

College of Computer Science, Sichuan University, Chengdu 610065, China



ARTICLE INFO

Article history:

Received 26 October 2021

Received in revised form 12 January 2022

Accepted 20 January 2022

Available online 25 January 2022

Keywords:

Person re-identification

Deep metric learning

Local feature learning

Generative adversarial learning

Sequence feature learning

ABSTRACT

In recent years, with the increasing demand for public safety and the rapid development of intelligent surveillance networks, person re-identification (Re-ID) has become one of the hot research topics in the computer vision field. The main research goal of person Re-ID is to retrieve persons with the same identity from different cameras. However, traditional person Re-ID methods require manual marking of person targets, which consumes a lot of labor cost. With the widespread application of deep neural networks, many deep learning-based person Re-ID methods have emerged. Therefore, this paper is to facilitate researchers to understand the latest research results and the future trends in the field. Firstly, we summarize the studies of several recently published person Re-ID surveys and complement the latest research methods to systematically classify deep learning-based person Re-ID methods. Secondly, we propose a multi-dimensional taxonomy that classifies current deep learning-based person Re-ID methods into four categories according to metric and representation learning, including methods for deep metric learning, local feature learning, generative adversarial learning and sequence feature learning. Furthermore, we subdivide the above four categories according to their methodologies and motivations, discussing the advantages and limitations of part subcategories. Finally, we discuss some challenges and possible research directions for person Re-ID.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, with the rapid development of intelligent surveillance devices and the increasing demand for public safety, a large number of cameras have been deployed in public places such as airports, communities, streets and campuses. These camera networks typically span large geographic areas with non-overlapping coverage and generate a large amount of surveillance video every day. We use this video data to analyze the activity patterns and behavioral characteristics of pedestrians in the real world for applications such as target detection, multi-camera target tracking and crowd behavior analysis. Person Re-ID can be traced back to the problem of multi-target multi-camera tracking (MTMCT tracking) [1], which aims to determine whether pedestrians captured by different cameras or pedestrian images from different video clips of the same camera are the same pedestrian [2]. Fig. 1 illustrates an example of a surveillance area monitored by multiple cameras with non-overlapping fields of view.

Fig. 2 shows the complete flow of the person Re-ID system, which mainly consists of two stages: pedestrian detection and re-

identification [3]. For pedestrian detection, many algorithms with high detection accuracy have emerged, such as YOLO [4], SSD [5] and Fast R-CNN [6]. Person Re-ID constructs a large image dataset (Gallery) from the detected pedestrian images and retrieves matching pedestrian images from it using probe images (Probe), so person Re-ID can also be regarded as an image retrieval task [7]. The key of person Re-ID is to learn discriminative features of pedestrians to distinguish between pedestrian images with the same identity and those with different identities. However, the difficulty of learning discriminative features of pedestrians is increased by the variation of view, pose, illumination and resolution in different cameras in the real world where pedestrians may appear in multiple cameras in multiple regions.

Traditional person Re-ID methods mainly used manual extraction of fixed discriminative features [8–12] or learned better similarity measures [13–16], which were error-prone and time-consuming, and greatly affected the accuracy and real-time performance of pedestrian Re-ID tasks. In 2014, deep learning was first used in the person Re-ID field [17,18]. Fig. 3 illustrates that there has been a significant increase in the proportion of collected person Re-ID papers over the years. Some researchers designed different loss functions to optimize the learning of discriminative features by network models [19–23]. Other researchers extracted more robust features of pedestrians by

* Corresponding author.

E-mail address: zhumin@scu.edu.cn (M. Zhu).

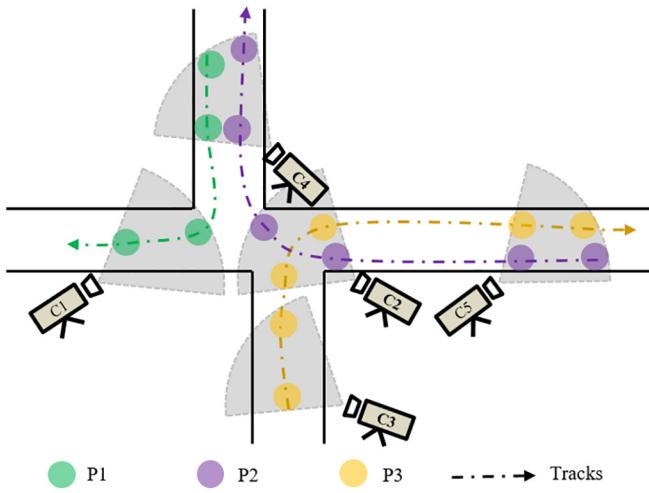


Fig. 1. Multi-camera surveillance network illustration of person Re-ID.

introducing local feature learning [24–28] or using attention mechanisms to focus on key information of body parts [29–34]. Ngo et al. [35,36] explored the method of high-level feature extraction aimed to explore context-based concept fusion by modeling inter-concept relationships, which were not modeled based on semantic reasoning [37]. Several works enhanced the final feature representation by combining global and local features of pedestrians [38–44]. Due to the good performance of GAN in generating images and learning features, generative adversarial learning was widely used for person Re-ID tasks [45–55]. To alleviate the shortage of information in single-frame images, some researchers used the complementary spatial and temporal cues of video sequences to effectively fuse more information in the video sequences [56–61]. Recently, graph convolutional network-based methods [60,62–65] also emerged to learn more discriminative and robust features by modeling graph relationships on pedestrian images. Some researchers [66,67] improved the robustness of the person Re-ID model by exploiting the information of a persons 3D shape. These methods are numerous and have different emphases. To give researchers a quick overview of the current state of development and valuable research directions in the field of person Re-ID, we conduct an in-depth survey of deep learning-based person Re-ID methods and summarize the relevant research results in recent years.

Prior to this survey, some researchers [3,68–77,77–81] also reviewed the person Re-ID field. In Table 1, we summarize the major contributions of these reviews. Some of these surveys [3,69] summarized image-based and video-based person Re-ID methods. Other surveys [70,73,74,77–79] summarized the deep learning-based person Re-ID methods in different dimensions, which developed rapidly after 2014 and became the main research means. Recently, Wang et al. [81] outlined methods of cross-domain person Re-ID and compared the

performance of these methods on public datasets. Yaghoubi et al. [80] proposed a multi-dimensional taxonomy to categorize the most relevant researches according to different perspectives. Zhou et al. [82] provided a review to summarize the developments in domain generalization for computer vision over the past decade. Behera et al. [83] reviewed traditional and deep learning person Re-ID methods in both contextual and non-contextual dimensions. Wu et al. [84] proposed new taxonomies for the two components of feature extraction and metric learning on person Re-ID. Behera et al. [85] conceptualized an overview of interpreting various futuristic cues on the IoT platform for achieving person Re-ID.

However, there are still some improvements to be made in these surveys, which lack the systematic classification and analysis of deep learning-based person Re-ID methods, also miss many discussions parts for person Re-ID. In this paper, compared to the above review, we focus more on metric learning and representation learning of deep learning methods in person Re-ID tasks and complement the latest research methods of recent years. We present an in-depth and comprehensive review of existing deep learning-based methods and discuss their advantages and limitations. We classify deep learning-based person Re-ID methods in terms of metric and representation learning dimensions, including four categories: deep metric learning, local feature learning, generative adversarial learning and sequence feature learning.

Deep metric learning focused on designing better loss functions for model training. Common loss functions for person Re-ID included: classification loss, verification loss, contrastive loss, triplet loss and quadruplet loss. Representation learning focused on developing feature construction strategies [78,86]. Therefore, we discussed the common feature learning strategies in recent person Re-ID methods, which were mainly in three categories: 1) Local feature learning, it learned part-level local features to formulate a combined representation for each person image; 2) Generative adversarial learning, it learned the image specific style representation or disentangled representation to achieve image-image style transfer or extract invariant features; 3) Sequence feature learning, it learned video sequence representation using multiple image frames and temporal information. In addition, we subdivided the above four categories based on their methodologies and motivations. This classification has a clear structure, which comprehensively reflects the most common deep metric learning methods and various representation learning methods in Re-ID tasks. Therefore, it is suitable for researchers to explore person Re-ID for practical needs. Furthermore, we attempt to discuss several challenges and research directions.

Specifically, the main contributions of our work are summarized as follows:

- We summarize the studies of several recently published person Re-ID surveys and complement the latest research methods to systematically classify deep learning-based person Re-ID methods.
- We comprehensively review the research methods of recent deep learning-based person Re-ID. Then, we propose a multi-dimensional

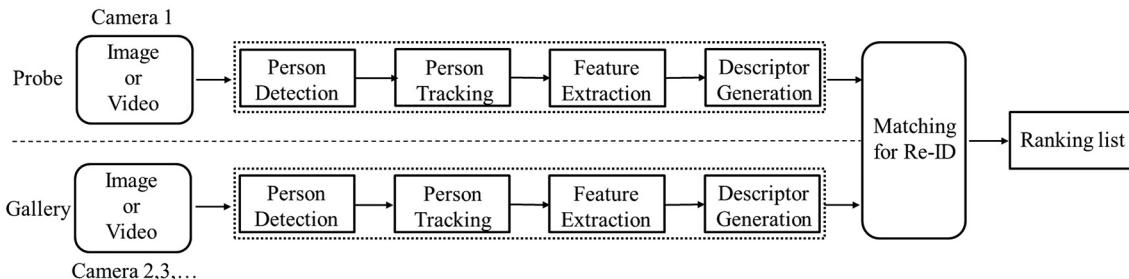


Fig. 2. Flowchart of person re-identification system.

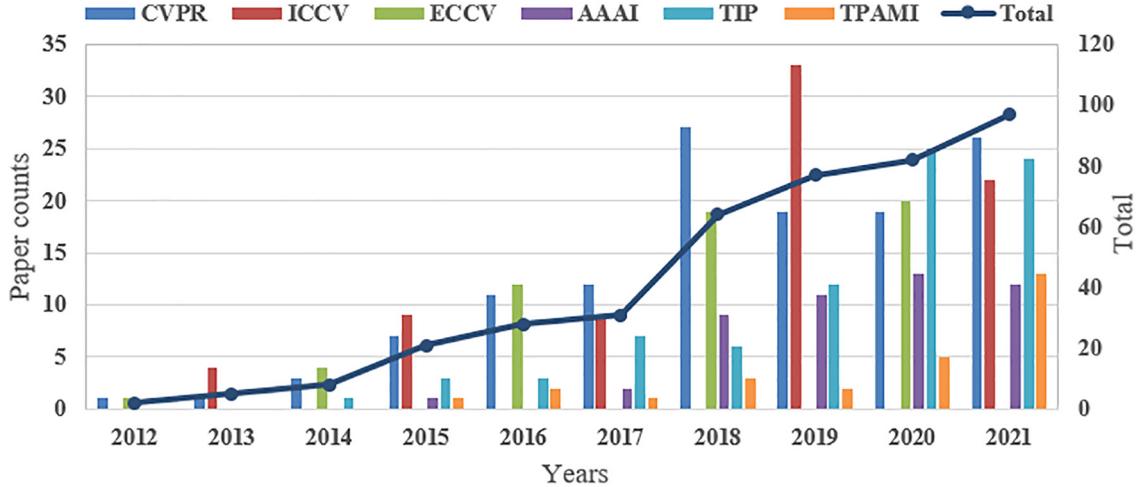


Fig. 3. The number of person Re-ID papers on top conferences and journals over the years.

taxonomy that classifies these methods into four categories according to metric and representation learning, including deep metric learning, local feature learning, generative adversarial learning, and sequence feature learning.

- We subdivide the above four categories based on their methodologies and motivations, discussing the advantages and limitations of part subcategories. This classification is more suitable for researchers to explore these methods from their practical needs.
- We summarize some existing challenges in the person Re-ID field and consider that there is still enough necessity to research it. Furthermore, we discuss seven possible research directions for person Re-ID researchers.

The remaining parts of this survey are structured as follows. In section 2, we discuss the common datasets and evaluation metrics used for person Re-ID benchmarks. In sec:Methods, we comprehensively review current deep learning-based methods for person Re-ID and divide them into four categories according to metric learning and representation learning. Furthermore, we subdivide the above four categories based on their methodologies and motivations, discussing the advantages and limitations of part subcategories. Finally, in sec:ConclusionAndFutureDirections, we summarize this paper and discuss the current challenges and future directions in the person Re-ID field.

2. Datasets and evaluation metrics

In this section, we present common datasets for evaluating existing deep learning-based person Re-ID methods in both image and video dimensions. In addition, we briefly describe common evaluation metrics for person Re-ID.

2.1. Datasets

In recent years, many methods have emerged to improve the performance of person Re-ID. However, the uncertainty of the real world brings about problems including occlusion, lighting changes, camera view switching, pose changes and similar clothing still cannot be well solved. These challenges make many algorithms still not available for real-world applications. Therefore, it is crucial to explore large-scale person Re-ID datasets covering more real scenes. As deep learning-based feature extraction methods gradually replace traditional manual feature extraction methods, deep neural networks require a large amount of training data, which has led to the rapid development of

large-scale datasets. The types of datasets and annotation methods vary greatly between datasets. Usually, the datasets used for person Re-ID can be divided into two categories, namely image-based and video-based datasets. The following subsection will introduce two types of commonly used datasets.

2.1.1. Image-based person Re-ID datasets

VIPeR [87] dataset is the first proposed small person Re-ID dataset. The VIPeR contains two viewpoint cameras, each of which captures only one image. The VIPeR uses manually labeled pedestrians and contains 1264 images for a total of 632 different pedestrians. Each image is cropped and scaled to a size of 128x48. The VIPeR dataset, which features multiple views, poses, and lighting variations, has been tested by many researchers, but it remains one of the most challenging person Re-ID datasets.

CUHK01 [88] dataset has 971 persons and 3884 manually cropped images, with each person also having at least two images captured in two disjoint camera views. In the CUHK01 dataset, camera A has more variations of viewpoints and poses, and camera B mainly includes images of the frontal view and the back view.

CUHK02 [89] dataset has 1816 persons and 7264 manually cropped images. The CUHK02 contains five pairs of camera views(ten camera views)), with each person also having at least two images captured in two disjoint camera views. Compared to the CUHK01 [88] dataset, the CUHK02 dataset has many identities and camera views and can obtain more configurations(which are the combinations of viewpoints, poses, image resolutions, lightings and photometric settings) of pedestrian images.

CUHK03 [14] dataset belongs to the large-scale person Re-ID dataset and is collected at the Chinese University of Hong Kong. The CUHK03 is acquired by 10 (5 pairs) cameras and provided with a manual marker and a deformable part model (DPM) detector [90] together to detect pedestrian bounding boxes. The CUHK03 contains 1360 different pedestrians with a total of 13,164 images, each of variable size. The CUHK03 improves on the CUHK01 [88] and CUHK02 [89] by increasing the number of cameras and captured images, thus capturing pedestrian images from more viewpoints. The CUHK03 dataset uses the pedestrian detection algorithm DPM to annotate pedestrians, making it more compatible with person Re-ID in the real world than using individual manual annotations.

Market-1501 [91] dataset is a large-scale person Re-ID dataset published in 2015, which was acquired using five high-resolution cameras and one low-resolution camera in front of a supermarket at Tsinghua University. The Market-1501 uses a pedestrian detector DPM to

Table 1

Comparison of several person Re-ID surveys in recent years.

Survey	Reference	Contribution
A survey of approaches and trends in person re-identification [68]	IVC14	1. Explored the problem of person Re-ID including system-level challenges, descriptor issues and correspondence issues; 2. Summarized the person Re-ID methods before 2016 including Contextual methods, Non-contextual methods and Active methods.
Person Re-identification Past, Present and Future [3]	arXiv16	1. Introduced the history of person re-ID and its relationship between image classification and instance retrieval; 2. Surveyed a broad selection of the hand-crafted systems and the large-scale methods in both image- and video-based re-ID.
A study on deep convolutional neural network based approaches for person re-identification [69]	PRMI17	1. Introduced image-based and video-based person Re-ID methods; 2. Discussed some important but undeveloped issues and future research directions.
Survey on person re-identification based on deep learning [70]	CAAI18	1. Some deep learning-based person Re-ID methods such as CNN-based, GAN-based and Hybrid-based methods are presented; 2. Proposed the direction of further research.
A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets [71]	TPAMI19	1. Presented an extensive review and performance evaluation of single-shot and multi-shot Re-ID algorithms. 2. Introduced the most recent advances in both feature extraction and metric learning.
Beyond intra-modality discrepancy: A comprehensive survey of heterogeneous person re-identification [72]	arXiv19	1. Surveyed the models that have been widely employed in heterogeneous person Re-ID. 2. Considered four cross-modality application scenarios: Low-resolution (LR), Infrared (IR), Sketch, and Text.
Deep learning-based methods for person re-identification: A comprehensive review [73]	NC19	1. Reviewed six types of methods of person Re-ID based on deep learning, including identification deep model, verification deep model, distance metric-based deep model, part-based deep model, video-based deep model and data augmentation-based deep model.
A Brief Survey of Deep Learning Techniques for Person Re-identification [74]	ICETCE20	1. Presented the issues of person Re-ID and the approaches used to solve these issues.
A Survey of Open-World Person Re-Identification [75]	TCSV20	1. Analyzed the discrepancies between closed-world and open-world scenarios. 2. Described the developments of open-set Re-ID works and their limitations.
Person search: New paradigm of person re-identification: A survey and outlook of recent works [76]	IVC20	1. Discussed about feature representation learning and deep metric learning with novel loss functions.
Survey on Reliable Deep Learning-Based Person Re-Identification Models: Are We There Yet? [77]	arXiv20	1. Surveyed state-of-the-art DNN models being used for person Re-ID task. 2. Discussed their limitations that can work as guidelines for future research.
Person re-identification based on metric learning: a survey [77]	MTA21	1. Summarized the research progress of person Re-ID methods based on metric learning.
Deep Learning for Person Re-identification: A Survey and Outlook [78]	TPAMI21	1. Reviewed for closed-world person Re-ID from three different perspectives, including deep

Table 1 (continued)

Survey	Reference	Contribution
		feature representation learning, deep metric learning and ranking optimization.
Survey on Unsupervised Techniques for Person Re-Identification [79]	CDS21	1. Reviewed the state-of-the-art unsupervised techniques of person Re-ID.
SSS-PR: A short survey of surveys in person re-identification [80]	PRL21	1. Proposed a multi-dimensional taxonomy to categorize the most relevant researches according to different perspectives. 2. Filled the gap between the recently published surveys.
Cross-Domain Person Re-identification: A Review [81]	AIC21	1. Reviewed methods of cross-domain person Re-ID. 2. Compared the performance of these methods on public datasets.

automatically detect pedestrian bounding boxes. It contains 1501 different pedestrians with a total of 32,668 images, each with a size of 128x64. Compared with CUHK03, Market-1501 has more annotated images and contains 2793 + 500 k interfering factors, and it is closer to the real world.

DukeMTMC-reID [92] dataset belongs to a subset of the MTMCT dataset DukeMTMC [93]. The DukeMTMC-reID dataset is collected at Duke University using eight static HD cameras. It contains 16 522 training images (from 702 persons), 2228 query images (from other 702 persons), and a search gallery (Gallery) of 17,661 images.

MSMT17 [45] dataset is a large-scale person Re-ID dataset published in 2018 and is captured at the campus by fifteen cameras. The MSMT17 uses the pedestrian detector Faster R-CNN [94] to automatically detect pedestrian-labeled frames. It contains 4101 different pedestrian information with a total of 126,441 images, which is one of the large datasets of pedestrian and annotated images in the current person Re-ID task. The MSMT17 dataset can cover more scenes than earlier datasets and contains more views and significant lighting variations.

Airport [71] dataset is constructed using video data from the six cameras installed at the postcentral security checkpoint at a commercial airport within the United States. The Airport dataset consists of 9651 identities, 31,238 distractors, a total of 39,902 images, and each one is cropped and scaled to a size of 128x64. The Airport uses pre-detected bounding boxes generated using aggregated channel features(ACF) [95] detector, which can accurately reflect real-world person Re-ID issues.

Table 2 shows the details of the above datasets. Most of the earlier image-based person Re-ID datasets (VIPeR, CUHK01, CUHK02, CUHK03, Market-1501) have the following limitations: (1) covering a single scene; (2) short time span without significant illumination variations; (3) expensive manual annotation or outdated automatic annotation with DPM detection. The CUHK03 contains viewpoint variations, detection errors, occlusions images. The Market-1501 contains viewpoint variations, detection errors and low-resolution images. But their simulation of the real world is relatively weak. The DukeMTMC-reID dataset contains more challenging attributes include viewpoint variations, illumination variations, detection errors, occlusions, and background clutter. The MSMT17 dataset collects images captured by 15 cameras for both indoor and outdoor scenes. Therefore, it presents complex scene transformations and backgrounds. The videos cover a long time, thus presenting complex lighting variations. The Airport dataset contains a large number of annotated identities and bounding boxes. To our best knowledge, MSMT17 and Airport are the largest and most challenging public datasets for person Re-ID.

2.1.2. Video-based person Re-ID datasets

PRID2011 [96] dataset is a video-based person Re-ID dataset proposed in 2011 and is obtained from two non-overlapping camera

Table 2

Typical image-based person Re-ID datasets.

Data type	Datasets	Years	ID	Boxes	Cameras	Labeled	Evaluation
Image	ViPeR [87]	2007	632	1264	2	Handcrafted	CMC
	CUHK01 [88]	2012	971	3884	2	Handcrafted	CMC
	CUHK02 [89]	2013	1816	7264	10	Handcrafted	CMC
	CUHK03 [14]	2014	1360	13,164	10	DPM + Handcrafted	CMC + mAP
	Market-1501 [91]	2015	1501	32,217	6	DPM + Handcrafted	CMC + mAP
	DukeMTMC-reID [52]	2017	1812	36,441	8	Handcrafted	CMC + mAP
	MSMT17 [45]	2018	4101	126,441	15	Faster RCNN	CMC + mAP
	Airport [71]	2019	9651	39,902	6	ACF	CMC + mAP

acquisitions and contains a total of 24,541 images of 934 different pedestrians with manually labeled pedestrian bounding boxes. The resolution size of each image is 128–64.

iLIDS-VID [97] dataset is acquired by two cameras at the airport and contains 300 pedestrians with 600 tracks totaling 42,495 pedestrian images. The iLIDS-VID uses manual annotation of pedestrian bounding boxes.

MARS [98] dataset is the first large-scale video-based person re-identification dataset proposed in 2016, which contains 1261 different pedestrians and approximately 20,000 video sequences of pedestrian trajectories acquired from six different cameras. The DPM detector and the generalized maximum multi-cornered problem (GMMCP) tracker [100] were used for pedestrian detection and trajectory tracking of MARS, respectively. The MARS dataset contains 3248 interfering trajectories and is fixed with 631 and 630 different pedestrians to divide the training and test sets, respectively. It can be considered as an extension of Market-1501.

DukeMTMC-VideoReID [19] dataset belongs to a subset of the MTMCT dataset DukeMTMC [93] for video-based person Re-ID. This dataset contains 1812 different pedestrians, 4832 pedestrian trajectories totaling 815,420 images, in which 408 pedestrians as interference terms, 702 pedestrians for training, 702 pedestrians for testing.

LPW(Labeled Person in the Wild) [99] dataset is a largescale video sequence-based person Re-ID dataset that collects three different crowded scenes containing 2731 different pedestrians and 7694 pedestrian trajectories with more than 590,000 images. The LPW dataset is collected in crowded scenes and has more occlusions, providing more realistic and challenging benchmarks.

Table 3 shows the details of video-based person Re-ID datasets. PRID2011 and iLIDS-VID use only two cameras to capture video and are labeled with fewer identities. That means other identities are only single cameras frame segments and the lighting and shooting angles of these identities in this dataset may not change much. MARS and DukeMTMC-ViedeReID are large-scale video-based person Re-ID datasets. Their bounding boxes and tracks are automatically generated and contain several natural detection or tracking errors, and each tag may have multiple tracks. LPW is one of the most challenging video-based person Re-ID datasets available and closer to the real world, distinguishes from existing datasets in three aspects: more identities and tracks, automatically detected bounding boxes, and far more crowded scenes with a larger time span.

Fig. 4 shows some sample images of a partial person Re-ID dataset. We can see that with the development of large-scale person Re-ID datasets, the number of pedestrian IDs and the number of labeled

frames or trajectories in the datasets are increasing, and the scenarios covered by the datasets are getting richer. These datasets use a combination of deep learning detectors and manual annotation to detect pedestrian bounding boxes, making the latest datasets closer and closer to the real world, thus enhancing the robustness of person Re-ID models. In addition, almost all mainstream person Re-ID datasets are evaluated using mean average precision (mAP) and cumulative matching characteristics (CMC) curves for performance evaluation.

2.2. Evaluation metrics

The commonly used evaluation metrics of person Re-ID algorithms are cumulative matching characteristics (CMC) curves and mean average precision (mAP).

In pattern recognition systems, CMC curves are important evaluation metrics in the fields of face, fingerprint, iris detection and person Re-ID, which can comprehensively assess the merits of model algorithms. Furthermore, CMC curves are considered to be a comprehensive reflection of the performance of the person Re-ID classifier. Before calculating the CMC curves, the probability Acc_k that the top-k retrieved images (top-k) in the gallery contain the correct query result is obtained by ranking the similarity between the query target and the target image to be queried, which is calculated as follows:

$$Acc_k = \begin{cases} 1 & \text{if top\text{--}k rank gallery samples} \\ & \text{contain query identity.} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

CMC curves are calculated by adding up the Acc_k of each query image and dividing it by the total number of query images, which are usually expressed as Rank-k. For example, Rank-1 accuracy indicates the probability of correctly matching to the first target in the matching list.

A single evaluation metric often cannot comprehensively evaluate the comprehensive performance of the person Re-ID algorithm. The mAP can reflect the extent to which all images with correct queries are at the front of the result queue in the query results. Considering both the average precision (AP) and precision-recall curve (PR) of the query process [91], instead of just focusing on the hit rate, which can measure the performance of person Re-ID algorithms more comprehensively. The algorithm is usually necessary to be evaluated separately for CMC curves and mAP in person Re-ID tasks. Zheng et.al. [8] proposed the Re-ranking method, which can re-rank the query results and further improve the effectiveness of Rank-k and mAP accuracy.

Table 3

Typical video-based person Re-ID datasets.

Data type	Datasets	Years	ID	Tracks	Cameras	Labeled	Evaluation
Video	PRID-2011 [96]	2011	934	400	2	Handcrafted	CMC
	iLIDS-VID [97]	2014	300	600	2	Handcrafted	CMC
	MARS [98]	2016	1261	20,715	6	DPM + GMMCP	CMC + mAP
	DukeMTMC-V [19]	2018	1812	4832	8	DPM	CMC + mAP
	LPW [99]	2018	2731	7694	11	DPM + NN + Handcrafted	CMC + mAP



Fig. 4. Sampled person images of person Re-ID datasets.

3. Deep learning based Re-ID method

In this section, we classify deep learning-based person Re-ID methods into four categories with classification structure is shown in Fig. 5, including methods for depth metric learning, local feature learning, generative adversarial learning and sequences feature learning. In addition, we subdivide the above four categories according to their methodologies and motivations, discussing and comparing the advantages and limitations of part subcategories.

3.1. Deep metric learning

Deep metric learning (DML) is one of the metric learning (ML) methods that aims to learn the similarity or dissimilarity between two pedestrian objects. The main goal of DML is to learn a mapping from the original image to the feature embedding (FE) such that the same pedestrians have smaller distances using a distance function on the feature space and different pedestrians feature farther apart from each other [101,102]. With the rise of deep neural networks(DNNs), DML has been widely used in computational vision, such as face recognition,

image retrieval, and person Re-ID. DML is mainly used to constrain the learning of discriminative features by designing loss functions for network models [78]. In this paper, we focus on loss functions commonly used in person Re-ID tasks, including classification loss [19,20,47,51,103–105], verification loss [20,78,106,107], contrastive loss [21,108,109], triplet loss [22,110,111] and quadruplet loss [23]. An illustration of five loss functions is shown in Fig. 6. These deep metric learning methods enable models to learn discriminative features automatically, which can solve the problem of manually designing features that consume a lot of labor costs.

3.1.1. Classification loss

Zheng et al. [3,112] treated the training process of person Re-ID as a multi-classification task for images and proposed an ID-discriminative embedding (IDE) network. The IDE treats each pedestrian as a different class and uses the ID of the pedestrian as a classification label to train a deep neural network, so the classification loss is also called ID loss. The training network for classification loss usually inputs a picture and connects a fully connected layer (FC) for classification at the end of the network, and then maps the feature vectors of the image onto the

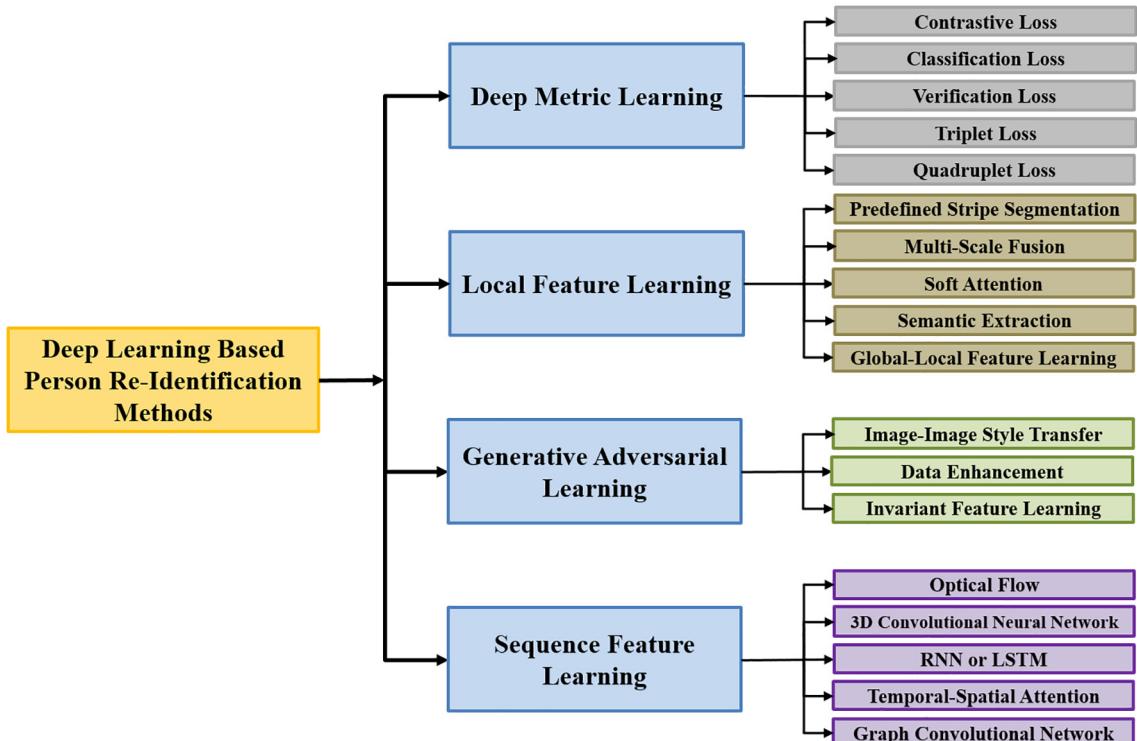


Fig. 5. Classification structure of deep learning-based person re-identification methods.

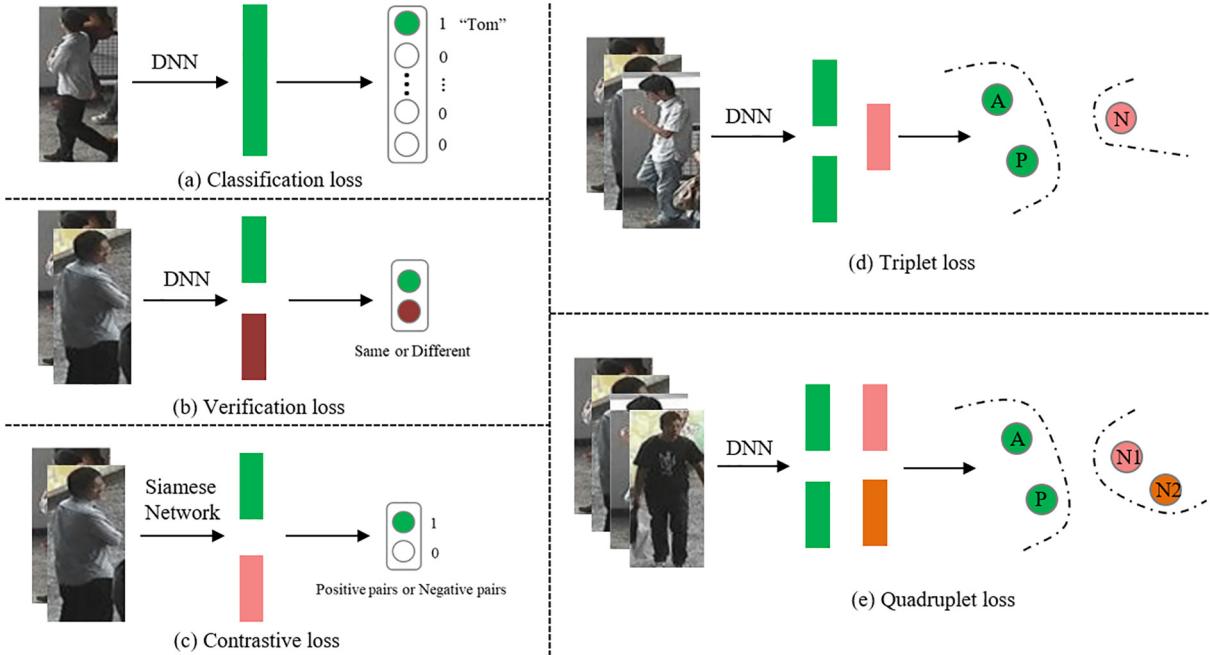


Fig. 6. An illustration shows a variety of existing deep metric learning-based person Re-ID losses. (a) Classification loss. (b) Verification loss. (c) Contrastive loss. (d) Triplet loss, where A , P , N indicate anchor, positive and negative samples, respectively. (e) Quadruplet loss, where $N1$ and $N2$ are different negative samples.

probability space by the softmax activation function. The cross-entropy loss for multi-classification of the person Re-ID task can be expressed as:

$$\mathcal{L}_{id} = - \sum_{a=1}^K q(x_a) \log p(y_a|x_a) \quad (2)$$

where K represents the number of training sample ID categories per batch, $q(x_a)$ denotes the label of sample image x_a . If x_a is identified as y_a , then $q(x_a) = 1$, otherwise $q(x_a) = 0$. $p(y_a|x_a)$ is the probability that picture x_a is predicted as category y_a using the softmax activation function. Classification loss is widely used as a depth metric learning for person Re-ID methods because of its advantages such as easy training of models and mining hard samples [19,20,47,51,103–105]. However, using ID information alone is not enough to learn a model with sufficient generalization ability. Therefore, ID loss usually needs to be combined with other losses to constrain the training of the model.

3.1.2. Verification loss

Person Re-ID can also be treated as a validation problem, and validation loss is proposed to guide the training of the model. In contrast to classification loss, the network trained by verification loss requires two images as input, and a binary loss is computed by fusing the feature information of the two images, which in turn determines whether the input two images are the same pedestrian [20,78,106,107]. The expression of the cross-entropy validation loss function is as follows:

$$\mathcal{L}_v = -y_{ab} \log p(y_{ab}|f_{ab}) - (1-y_{ab}) \log (1-p(y_{ab}|f_{ab})) \quad (3)$$

Supposing the network inputs two images x_a and x_b , we get the feature vectors f_a and f_b of these two images respectively, and calculate the difference feature $f_{ab} = (f_a - f_b)^2$ of the two feature vectors. We use the softmax activation function to calculate the probability p that the image pairs x_a and x_b have the same pedestrian ID, where y_{ab} is the pedestrian ID label of the two images. When the images x_a and x_b have the same ID, $y_{ab} = 1$, otherwise, $y_{ab} = 0$.

The verification loss is less efficient in recognition because it can only input a pair of images to judge the similarity when tested while

ignoring the relationship between the image pair and other images in the dataset. For this reason, researchers considered combining classification and validation networks [106,107], and the combined loss can be expressed as $\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_v$. The hybrid loss can combine the advantages of classification loss and verification loss, which can predict the identity ID of pedestrians and perform similarity metrics simultaneously, thus improving the accuracy of person Re-ID.

3.1.3. Contrastive loss

Contrastive loss, which mainly constrains the similarity or dissimilarity between pairs of data, is generally used for model training of twin networks (Siamese Network) in person Re-ID tasks [21,108,109]. Its function can be expressed as:

$$\mathcal{L}_c = yd(x_a - x_b)^2 + (1-y)[m - d(x_a - x_b)^2]_+ \quad (4)$$

where $[z]_+ = \max(0, z)$, x_a and x_b are two images input to the twin network at the same time. $d(x_a, x_b)$ usually indicates the euclidean distance (similarity) of the two images. m is the set training threshold, and y is the label of whether each pair of training images matches. When $y = 1$, it means that the input images x_a and x_b belong to the pedestrians with the same ID (positive sample pair). When $y = 0$, it means that the input images x_a and x_b belong to pedestrians with different IDs (negative sample pair). \mathcal{L}_c reflects well the matching degree of sample pairs, which is often used to train models for person Re-ID feature extraction and often works together with classification loss combinations for training networks [47].

3.1.4. Triplet loss

Triplet loss is one of the most widely used depth metric losses in person Re-ID tasks, and it aims to minimize the intra-class distance and maximize the inter-inter-class distance of samples. With the development of deep neural networks, a large number of variants based on triplet loss have emerged [22,110,111,113]. The triplet loss function can be expressed as:

$$\mathcal{L}_{trip} = [m + d(x_a, x_p) - d(x_a, x_n)]_+ \quad (5)$$

Different from contrast loss, the input of triplet loss is a triplet consisting of three images. Each triplet contains a pair of positive samples and a negative sample, where x_a is the Anchor image, x_p is the Positive image, and x_n is the Negative image, and the pedestrians of x_a and x_p have the same ID. The pedestrians of x_a and x_n have different IDs. By model training, the distance between x_a and x_p in the Euclidean space is made closer than the distance between x_n and x_a . To improve the performance of models, some deep learning-based person Re-ID methods use a combination of classification loss and triplet loss [114–118]. Experiments have shown that combining these two losses facilitates the model to learn discriminative features. Traditional triplet loss randomly selects three images from the training set during training, which may result in a simple combination of samples and lacks the training of hard sample combinations and makes the training model less generalizable. For this reason, some researchers considered improving triplet loss for mining hard samples [22,119,120].

3.1.5. Quadruplet loss

Another improvement for the triplet loss is to add a negative sample picture X_{n2} to form a quadruplet loss [23], where negative sample X_{n1} and negative sample X_{n2} have different pedestrian IDs. The expression of the quadruplet loss function is:

$$\mathcal{L}_{quad} = [m_1 + d(x_a, x_p) - d(x_a, x_{n1})]_+ + [m_2 + d(x_a, x_p) - d(x_{n1}, x_{n2})]_+ \quad (6)$$

where m_1 and m_2 are custom training thresholds. The positive and negative sample pairs have the same anchored image x_a . The first term of \mathcal{L}_{quad} is identical to the triplet loss function, which is used to constrain the relative distance between positive and negative sample pairs. The traditional triplet loss function often increases the inter-class distance of negatives sample pairs, which affects the feature learning of image x_a . For this reason, \mathcal{L}_{quad} introduces a second term to constrain the absolute distance between positive and negative sample pairs. The positive and negative sample pairs in the second term have different anchor images, which can effectively reduce the intra-class distance of positive sample pairs while increasing the distance of negative sample pairs between classes. In order to make the first term play a dominant role, it is usually important to ensure that $m_1 > m_2$ during the training process. However, most person Re-ID methods using triplet loss drive focused more on differentiating appearance differences and cannot effectively learn fine-grained features. To address this issue, Yan et al. [121] introduce a novel pairwise loss function that enables Re-ID models to learn the fine-grained features by adaptively enforcing an exponential penalization.

3.2. Local feature learning

Based on the features extracted from pedestrian images for classification, person Re-ID methods can be classified into global feature learning-based methods and local feature learning-based methods. The global feature learning methods usually extract one feature of the pedestrian image [122–124], and it is difficult for this method to capture the detailed information of the pedestrian. Therefore, how to extract discriminative local features of pedestrians with subtle differences becomes a problem for researchers to focus on.

The local feature learning-based methods aim at learning pedestrian discriminative features and ensuring the alignment of each local feature. Manual annotation or neural networks are usually used to automatically focus on certain local regions with key information and extract the distinguishing features from these regions. Commonly used local feature learning methods are predefined stripe segmentation [24,25,27,108,125,126], multi-scale fusion [127–131], soft attention

[26,29–34,132,133], pedestrian semantic extraction [27,28,31,132,133] and global-local feature learning [38–44]. These methods can alleviate the problems of occlusion, boundary detection errors, view, and pose variations.

3.2.1. Predefined stripe segmentation

The main idea of the method based on predefined stripe segmentation is to strip the learned features according to some predefined division rules, which must ensure that the partitions are spatially aligned. Liu et al. [125] proposed an attribute and appearance-based contextual attention network, where the appearance network learns spatial features from the whole body, horizontal and vertical parts of the pedestrian. Varior et al. [108] divided the pedestrian image into several strips uniformly and extracted local features from each strip image block. Sun et al. [24] considered the content consistency within each stripe to propose a local convolutional baseline (PCB). The PCB uses a uniform feature partitioning strategy to learn local features and outputs convolutional features consisting of multiple stripes to enhance the consistency of each partition's feature content, thus ensuring stripes are spatially aligned.

Although the above methods can extract discriminative features for the striped areas, it may lead to incorrect retrieval results as the model is unable to distinguish between the obscured and unobscured areas. To relieve occlusion, Sun et al. [25] proposed a visibility-aware local model based on PCB to ensure that local features are spatially aligned and avoid interference due to pedestrian occlusion by learning common region features that are visible in both images. Fu et al. [126] horizontally slice the deep feature maps into multiple spatial strips using various pyramid scales and used global average pooling and maximum pooling to obtain discriminative features for each strip, which was named horizontal pyramid pooling (HPP). HPP can ignore that interference information, mainly coming from similar clothing or background.

3.2.2. Multi-scale fusion

Small-scale feature maps have a strong ability to represent spatial geometric information and can obtain detailed information of the image. Large-scale feature maps are good at characterizing semantic information and can get contour information of images. Extracting pedestrian features at multiple scales for fusion can obtain rich pedestrian feature representation.

Liu et al. [127] proposed a multi-scale triple convolutional neural network, which can capture pedestrian appearance features at different scales. Since pedestrian features learned at different scales differ or conflict, the direct merging of features at multiple scales may not achieve the best fusion effect. Therefore, researchers have started to focus on the complementary advantages of cross-scale implicit associations. Chen et al. [128] studied the problem of person Re-ID multi-scale feature learning and proposed a deep pyramidal feature learning deep neural network framework, which can overcome the differences in cross-scale feature learning while learning multi-scale complementary features. Zhou et al. [135] presented a Re-ID CNN termed Omni-scale network (OSNet) to learn features that not only captured different spatial scales but also encapsulated a synergistic combination of multiple scales. In traditional person Re-ID datasets, OSNet achieved state-of-the-art performance, despite being much smaller than existing Re-ID models. The challenge of large intra-class variation and small inter-class variation often arises in cross-camera person Re-ID tasks. For example, cross-camera viewpoint changes can obscure parts of the person with discriminative features, or pedestrians wearing similar clothes appear across cameras, which makes the matching of the same person incorrect.

3.2.3. Soft attention

The goal of attention is to find the areas that have a greater impact on the feature map and to focus the model on discriminative local parts of body appearances to correct misalignment, eliminate background

perturbation. Due to the good performance of the attention mechanism in the computer vision field, it is often used as a local feature learning in the person Re-ID tasks. Most of the current attention-based person Re-ID methods tend to use soft attention which can be divided into spatial attention, channel attention, mixed attention, non-local attention and position attention. Liu et al. [29] proposed an attention-based deep neural network capable of capturing multiple attention features from the underlying to the semantic layer to learn fine-grained integrated features of pedestrians. Li et al. [30] used balanced attention convolutional neural networks to maximize the complementary information of attention features at different scales to solve the person Re-ID challenge for arbitrary unaligned images. To obtain the local fine-grained features of a person, Ning et al. [136] proposed a multi-branch attention network with diversity loss, and the local features were obtained via adaptive filtering by removing interference information.

The above spatial attention-based methods tend to focus only on the local discriminative features of pedestrians but ignore the impact of feature diversity on pedestrian retrieval. Chen et al. [26] proposed an attentional diversity network that used complementary channel attention module (CAM) and position attention module (PAM) to learn the characteristics of pedestrian diversity. Considering that features extracted from first-order attention such as spatial attention and channel attention are not discriminative in complex camera view and pose change scenarios [32]. Chen et al. [33] proposed a higher-order attention module, which modelled the complex higher-order information in the attention mechanism to mine discriminative attention features among pedestrians.

3.2.4. Semantic extraction

Some researchers used deep neural networks to extract semantic information such as body parts or body postures instead of bounding boxes to extract local features from pedestrian body parts to improve the performance of person Re-ID. Zhao et al. [28] considered the application of body structure information to the person Re-ID task and proposed a novel CNN, called SpindleNet. Specifically, firstly, SpindleNet used a body part generation network to locate 14 key points of body parts to extract 7 body regions of the pedestrian. Secondly, SpindleNet captured semantic features from different body regions using a convolutional neural network. Finally, SpindleNet used a tree fusion network with competing strategies to merge the semantic features from different body regions. SpindleNet can align the features of body parts over the whole image and can better highlight local detail information.

Sometimes, not only do the body parts contain discriminative features but also non-body parts may contain certain key features, such as the pedestrian's distinguishing backpack or handbag. Therefore, some researchers considered the alignment of non-body parts. Guo et al. [114] proposed a dual part-aligned representation scheme that captured distinguishing information beyond body parts using an attention mechanism to update the representation by exploiting the complementary information from both the accurate human parts and the coarse non-human parts. Miao et al. [27] proposed a pose-guided feature alignment scheme to distinguish information from occlusion noise by pedestrian pose bounding markers, thus aligning the query image and the non-occluded area of the queried image.

3.2.5. Global-local feature learning

Local feature learning can capture detailed information about a region of the pedestrian, but the reliability of local features can be affected by variations in pose and occlusion. Therefore, some researchers often combine fine-grained local features with coarse-grained global features to enhance the final feature representation. Wang et al. [38] proposed a multi-granularity feature learning strategy with global and local information, including one branch for global feature learning and two branches for local feature learning. Ming et al. [137] designed a global-local dynamic feature alignment network (GLDFA-Net)

framework, which contained both global and local branches. The local sliding alignment(LSA) strategy was introduced into the local branch of GLDFA-Net to guide the computation of distance metrics, which can further improve the accuracy of the testing phase. To mitigate the impact of imprecise bounding boxes on pedestrian matching, Zheng et al. [43] proposed a coarse-grained to fine-grained pyramid model that integrates not only local and global information of pedestrians but also progressive cues from coarse to fine-grained. The model can match pedestrian images of different scales and retrieve pedestrian images with the same local identity even when the images are not aligned.

3.2.6. Comparison and discussion

Table 4 shows the experimental results of the local feature learning methods on CUHK03, Market1501 and DukeMTMC-reID datasets. These results are all experimental results without Re-ranking [8]. In a general view, the experimental performance of the semantic extraction and global-local feature learning methods is significantly higher than that of the methods with predefined stripe segmentation, multi-scale fusion and part attention.

In general, the predefined stripe segmentation method is simple and easy to implement, but it is hard segmentation and requires high image alignment. With the change of real scene camera view and pedestrian pose, the hard segmentation strategy cannot solve the problem of unaligned pedestrians well. The multi-scale fusion method can learn the deeper cues of pedestrian images, but there will be redundancy and conflicting features at different scales. The attention focuses only on the local features of key parts of pedestrians, and easily ignores the distinguishing features of non-focus regions. The semantic extraction method can precisely locate the local features of pedestrians by learning the structural information of pedestrian pose, but it requires the additional computation of pedestrian pose models. The global-local feature learning method can effectively utilize the complementary advantages of global features and local features and is one of the common methods used by researchers to improve model performance.

3.3. Generative adversarial learning

In 2014, Goodfellow et al. [139] first proposed generative adversarial networks (GAN) and rapidly developed in recent years. A large number of variants and applications of GAN emerged [45,47,51,54,55,140–143]. Image generation, as one of the significant applications of GAN, was widely used in the field of person Re-ID. **Fig. 7** shows the workflow diagram of GAN used to generate the image. In the training phase, the generator G_{AB} converts image A into image B with random noise, the generator G_{BA} converts image B into image A , and the discriminator D_B determines whether the generated image B approximates the original image B style (Real or Fake). The generator and discriminator keep adversarial until convergence by minimizing the discriminator loss and L_2 loss [2].

Some researchers used GAN to transform the style of images or unify different image styles to mitigate image style differences between various datasets or within the same dataset [45,47–51,144–147]. Some works used GAN to synthesize pedestrian images with a different pose, appearance, lighting, and resolution for expanding the dataset to improve the generalization ability of the model [52–54,140,148–153]. Some researchers also used GAN to learn features that are not noise-related but identity-related to improve the accuracy of feature matching [46,55,141,154]. These methods can alleviate the small number of training samples, resolution, illumination, view, and pose variation. Based on the characteristics and application scenarios of GAN, we classify the generative adversarial learning-based person Re-ID methods into three categories: image-image style transfer, data enhancement, and invariant feature learning. For image-image style transfer methods, GAN learned the background, resolution, lighting and other features of an image and transferred these features to other images to give other images a different style. For data enhancement methods, the diversity of

Table 4

Comparison of experimental results of local feature learning based methods. * represents the use of multiple deep learning methods. The bold and underlined numbers represent the top two results.

Categories	Methods	CUHK03		Market-1501		DukeMTMC-reID		Reference
		mAP	R-1	mAP	R-1	mAP	R-1	
Predefined stripe segmentation	PCB + RPP [24]	–	–	81.6	93.8	69.2	83.3	ECCV'18
	CA3Net [125]	–	–	80.0	93.2	70.2	84.6	ACMMM'18
	PGFA [27]*	–	–	76.8	91.2	65.5	82.6	ICCV'19
	VPM [25]	–	–	80.8	93.0	72.6	83.6	CVPR'19
	HPM [126]	57.5	63.9	82.7	94.2	74.3	86.6	AAAI'19
Multi-scale fusion	DPFL [128]	40.5	43.0	72.6	88.6	60.6	79.2	ICCV'17
	MSCAN [130]	–	74.2	66.7	86.8	–	–	CVPR'17
	OSNet [129]	67.8	72.3	84.9	94.8	73.5	88.6	ICCV'19
	CAM [131]	64.2	66.6	84.5	94.7	72.9	85.8	CVPR'19
Soft attention	HydraPlus [29]	–	91.8	–	76.9	–	–	ICCV'17
	HA-CNN [30]	44.4	41.0	75.7	91.2	63.8	80.5	CVPR'18
	ABD-net [26]	–	–	88.3	95.6	78.6	89.0	ICCV'19
	HOA [33]	72.4	77.2	85.0	95.1	77.2	89.1	ICCV'19
Semantic extraction	AANet [34]	–	–	83.4	93.9	74.3	87.7	CVPR'19
	HLGAT [132]	80.6	83.5	93.4	97.5	87.3	92.7	CVPR'21
	PAT [138]	–	–	88.0	95.4	78.2	88.8	CVPR'21
	Spindle [28]	–	88.5	–	76.9	–	–	CVPR'17
Global-Local feature learning	SPReID [134]	–	94.3	83.4	93.7	73.3	85.9	CVPR'18
	P2-Net [114]	73.6	78.3	85.6	95.2	73.1	86.5	ICCV'19
	GLAD [39]	–	85.0	73.9	89.9	–	–	ACMMM'17
	PDC [40]	–	88.7	63.4	84.1	–	–	ICCV'17
Semantic extraction	Pyramid [43]	76.9	78.9	88.2	95.7	79.0	89.0	CVPR'19
	RGA [42]	77.4	81.1	88.4	96.1	–	–	CVPR'20
	SCSN [41]	84.0	86.8	88.5	95.7	79.0	91.0	CVPR'20

samples that can be generated by GAN to expand the dataset was used to reinforce the final feature representation. For invariant feature learning, GAN was used for disentangled representation learning, which can learn identity-related but noise-independent features (e.g., pose, lighting, resolution, etc.).

3.3.1. Image-image style transfer

Domain gaps usually exist between different datasets in the person Re-ID task [45]. When training and testing on various datasets separately, the performance of the model is severely degraded, which hinders the effective generalization of the model to new test sets [155,156]. A common strategy to solve such as domain gap is to use GAN to perform style transformation across data domains. Since CycleGAN [142] implemented conversion of any two image styles, researchers considered improving on this to achieve adaptive pedestrian style conversion between different datasets to reduce or eliminate domain aberrations. Inspired by CycleGAN, Wei et al. [45] proposed a

person transfer generative adversarial network (PTGAN) to transfer the pedestrians in the source domain to the target dataset while preserving the identity of the pedestrians in the source domain so that the pedestrians in the source domain have the background and lighting styles of the target domain. The pedestrians in the source domain are transferred to the target dataset so that the pedestrians in the source domain have the background and lighting patterns of the target domain.

Deng et al. [47] used twin networks and CycleGAN to form a similarity preserving generative adversarial network (SPGAN) to migrate labeled pedestrians from the source domain to the target domain in an unsupervised manner. Liu et al. [48] proposed an adaptive transfer network (ATNet). ATNet used three CycleGANs to implement the style of camera view, lighting, resolution and adaptively assign weights to each CycleGAN according to the degree of influence of different factors. Zhong et al. [50] proposed hetero-homogeneous learning (HHL) method that considers not only the domain differences between various datasets but also the effect of style differences of

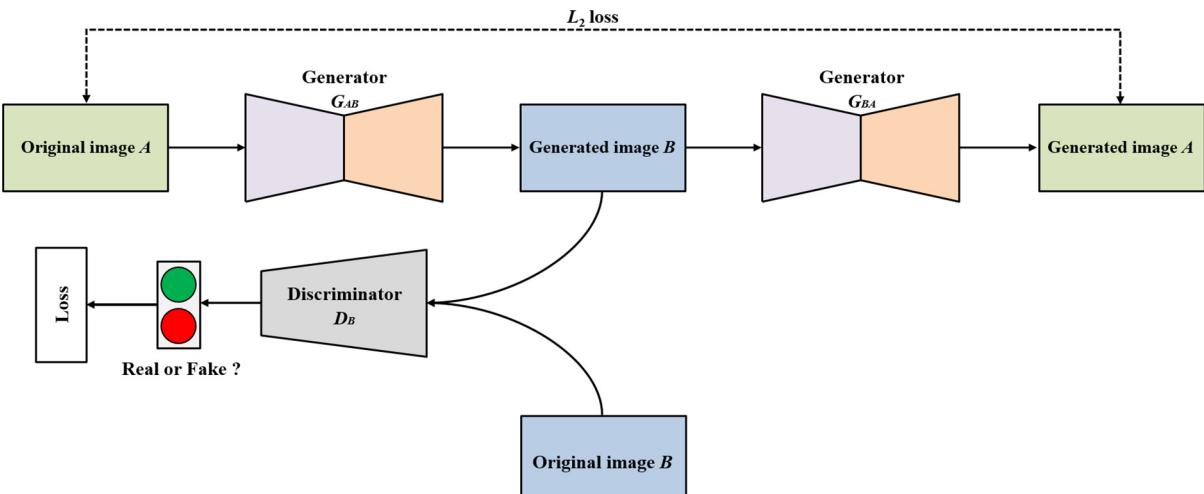


Fig. 7. Workflow diagram of GAN for image generation.

cameras within the target domain on the cross-domain adaptation person Re-ID performance.

Zhong et al. [51] introduced camera style(CamStyle) to solve the problem of style variation among different cameras within the same dataset. CamStyle used CycleGAN to migrate the labeled training data to various cameras so that the synthesized samples had the styles of different cameras while retaining the pedestrian labels. In addition, CamStyle can also smooth out style differences between various cameras in the same dataset.

3.3.2. Data enhancement

Unlike style transformation that uses GAN to reduce domain gaps, the data augmentation-based methods start with the training of the model and improve the generalization ability of the model by increasing the diversity of the training data. Zheng et al. [52] were the first to use deep convolutional generative adversarial networks (DCGAN) [157] to generate sample data. Huang et al. [53] proposed a multi-pseudo regularized label (MpRL), which assigned an appropriate virtual label to each generated sample to establish the correspondence between the real image and the generated image. MpRL effectively distinguished various generated data and achieved good recognition results on datasets such as Market-1501, DukeMTMC-Reid and CUHK03.

Liu et al. [140] introduced pedestrian pose information to assist GAN in generating samples. The GAN was used to generate sample images with both the pose structure of pedestrians in MARS data and the appearance of pedestrians in the existing dataset. Qian et al. [150] used pose normalization GAN (PN-GAN) to generate pedestrian images with uniform body pose. To alleviate the problem that the pose of pedestrian images generated by earlier methods is prone to large deviations, Zhu et al. [151] trained a discriminator using a multilayer cascaded attention network. The discriminator can efficiently optimize the pose transformation of pedestrians using pose and appearance features so that the generated pedestrian images have the better pose and appearance consistency with the input images.

3.3.3. Invariant feature learning

GAN can be used for feature learning in addition to synthesizing images. Generally, the person Re-ID task in the real-world consists of high-level and low-level vision variations [46]. The former mainly includes changes in pedestrian occlusion, pose, and camera view, and the latter mainly includes changes in resolution, illumination, and weather. The images obtained from low-level vision changes are usually called degraded images. These vision changes may lead to the loss of discriminative feature information, which may cause feature mismatch and significantly degrade the retrieval performance [158].

For pose changes in high-level vision, Ge et al. [55] proposed feature distilling generative adversarial network (FD-GAN) to learn features related to pedestrian identity instead of pose for pedestrians. The method requires no additional computational cost or auxiliary attitude

information and has advanced experimental results on the Market-1501, CUHK03 and DukeMTMC-reID.

Several researchers considered using GAN to learn common invariant features of low- and high-resolution pedestrian images. Chen et al. [141] proposed an end-to-end resolution adaptation and re-identification network (RAIN) that learned and aligned invariant features of pedestrian images of different resolutions by adding adversarial losses to low- and high-resolution image features. Li et al. [154] proposed adversarial learning strategies for cross resolution, which not only learned invariant features of pedestrian images of different resolutions but also recovered the lost fine-grained detail information of low-resolution images using super-resolution (SR).

3.3.4. Comparison and discussion

Table 5 shows the experimental results of GAN-based methods on CUHK03, Market1501 and DukeMTMC-reID datasets. These results are all experimental results without Re-ranking. Both IDE [3] + CameraStyle [51] and IDE [3] + UnityGAN [145] methods use GAN to generate pedestrian images with different camera styles within the same dataset with great experimental performance obtained in the Market-1501 and DukeMTMC-reID datasets. FD-GAN [55] has the highest performance in CUHK03, and it learns features related to pedestrian identity and poses independently, effectively reducing the influence of pose on the accuracy of person Re-ID.

We use the Market-1501 (M) and DukeMTMC-reID (D) datasets as the source and target domains, respectively. **Table 6** compares traditional manual feature extraction methods (LOMO [159] and Bow [91]), traditional unsupervised methods (UMDL [160], CAMEL [16] and PUL [161]), and GAN-based cross-domain style transfer methods (CycleGAN(base) [142], PTGAN [47], SPGAN [47], HHL [50], ATNet [48] and CR-GAN [49]). From the experimental results, the cross-domain style transformation method is significantly better than the traditional unsupervised learning and manual feature learning methods.

In general, the method of image-to-image style transformation smooths the style variation of pedestrian images in different domains. Such methods can obtain a large number of automatically labeled synthetic images with the target domain style, which can be used together with the original images to enhance the training set and mitigate the domain gaps between different datasets. The problem with these methods is that the synthetic images contain noise, which may conflict with the source domain images when used for model training and affect the learning of discriminative features by the model. The method of generating diverse pedestrian images using GAN alleviates the problem of insufficient available training data to a certain extent. The method of image synthesis without auxiliary information guidance cannot generate high-quality images with sufficient distinguishing information. Auxiliary information-guided image synthesis methods require complex network structures to learn various pedestrian poses, which adds additional training costs. Invariant feature learning methods can alleviate

Table 5

Comparison of experimental results of different GAN-based methods on CUHK03, Market1501 and DukeMTMC-reID datasets. The bold and underlined numbers represent the top two results.

Categories	Methods	CUHK03		Market-1501		DukeMTMC-reID		Reference
		mAP	R-1	mAP	R-1	mAP	R-1	
Image-image style transfer	IDE [3] + CameraStyle [51]	–	–	68.7	88.1	57.6	78.3	CVPR'18
	IDE [3] + UnityStyle [145]	–	–	89.3	<u>93.2</u>	65.2	82.1	CVPR'20
Data enhancement	LSRO [52]	87.4	84.6	66.1	84.0	47.1	67.7	ICCV'17
	PNGAN [150]	–	79.8	72.6	89.4	53.2	73.6	ECCV'18
Invariant feature learning	PT [140]	42.0	45.1	68.9	87.6	56.9	78.5	CVPR'18
	MpRL [53]	<u>87.5</u>	85.4	67.5	85.8	58.6	78.8	TIP'19
	DG-Net [54]	–	–	86.0	94.8	74.8	86.6	CVPR'19
	FD-GAN [55]	91.3	92.6	<u>77.7</u>	90.5	64.5	80.0	NIPS'18
	RAIN [141]	–	78.9	–	–	–	–	AAAI'19
	CAD-Net [154]	–	82.1	–	83.7	–	75.6	ICCV'19
	DI-REID [46]	–	85.7	–	–	–	–	CVPR'20

Table 6

Comparison of experimental results of different GAN-based cross-domain methods. The bold and underlined numbers represent the top two results.

Methods	D M		M D		Reference
	mAP	R-1	mAP	R-1	
LOMO [159]	8.0	27.2	4.8	12.3	CVPR'15
Bow [91]	14.8	35.8	8.3	17.1	ICCV'15
UMDL [160]	12.4	34.5	7.3	18.5	CVPR'16
CAMEL [16]	26.3	54.5	—	—	ICCV'17
PUL [161]	20.5	45.5	16.4	30.0	MM'18
CycleGAN [142]	19.1	45.6	19.6	38.1	CVPR'17
PTGAN [45]	—	38.6	27.4	—	CVPR'18
SPGAN [47]	22.8	51.5	22.3	41.4	CVPR'18
HHL [50]	31.4	62.2	27.2	46.9	ECCV'18
ATNet [48]	25.6	55.7	24.9	45.1	CVPR'19
CR-GAN [49]	29.6	59.6	30.0	52.2	ICCV'19
DG-Net ++ [162]	61.7	82.1	63.8	78.9	ECCV'20
GCL [153]	75.4	90.5	67.6	81.9	CVPR'21

the problem of unaligned pedestrian features and improve the accuracy of person Re-ID by learning features related to pedestrian identity but not to pose, resolution and illumination.

3.4. Sequence feature learning

There have been many researchers who have used the rich information contained in video sequences for person Re-ID. These sequence feature learning-based methods take short videos as input and use both spatial and temporal complementary cues, which can alleviate the limitations of appearance-based features. Most of these methods use optical flow information [56,163–166], 3-dimensional convolutional neural networks (3DCNNs) [57,167], recurrent neural networks(RNN) or long short term memory(LSTM) [164,165,168,169], spatial-temporal attention [58,59,166,170–173] or graph convolutional networks (GCN) [60–62,174] to model the spatial-temporal information of video sequences. These methods can mitigate occlusions, resolution changes, illumination changes, view and pose variations, etc.

3.4.1. Optical flow

The optical flow method uses the change of pixels in the video sequence in the time domain and the correlation of the spatial-temporal context of adjacent frames to obtain the correspondence between the previous frame and the current frame. This method can obtain the motion information of the target between adjacent frames. Chung et al. [56] proposed a dual-stream convolutional neural network (DSCNN), where each stream is a siamese network. DSCNN can model both RGB images and optical flow, learning spatial and temporal information separately, allowing each siamese network to extract the best feature representation. Liu et al. [163] proposed an accumulative context network (AMOC), which consists of two input sequences, feeding the original RGB image and the optical flow image containing motion information, respectively. AMOC was used to improve the accuracy of person Re-ID by learning the discriminative cumulative motion context information of video sequences. The optical flow method was often used in combination with other methods such as McLaughlin et al. [165] who used optical flow information and RGB colors of images to capture motion and appearance information, combined with RNN to extract complete pedestrian appearance features of video sequences.

3.4.2. 3D convolutional neural network

Three-dimensional convolutional neural networks (3DC-NN) are capable of capturing temporal and spatial feature information in videos. Recently, some researchers have applied 3DCNN to video-based person Re-ID with good results. Liao et al. [57] proposed a video person Re-ID method based on a combination of 3DCNN and non-local attention. 3DCNN used 3D convolution on video sequences to extract aggregated representations of spatial and temporal features and used non-local

spatial-temporal attention to solve the alignment problem of deformed images. Although 3DCNN exhibited better performance, the stacked 3D convolution led to significant growth of parameters. Too many parameters not only made 3DCNN computationally expensive but also led to difficulties in model training and optimization. This made 3DCNN not readily applicable on video sequence-based person Re-ID, where the training set was commonly small and person ID annotation was expensive. To explore rich temporal cues for person Re-ID while mitigating the shortcomings of existing 3DCNN models, Li et al. [167] proposed a dual-stream multiscale 3D convolutional neural network (M3DCNN) for extracting spatial-temporal cues for video-based person Re-ID. M3DCNN was also more efficient and easier to optimize than the existing 3DCNN.

3.4.3. RNN or LSTM

RNNs or LSTM can extract temporal features and are often applied in video-based person Re-ID tasks. McLaughlin et al. [165] proposed a novel recursive convolutional network(RCN), which used CNN to extract spatial features of video frames and RNNs to extract temporal features of video sequences. Yan et al. [169] used a recurrent feature aggregation network based on LSTM, which obtained cumulative discriminative features from the first LSTM node to the deepest LSTM node and effectively alleviated interference caused by occlusion, background clutter and detection failure. Chen et al. [164] decomposed a video sequence into multiple segments and used LSTM to learn the segments where the probe images are located in temporal and spatial features. This method reduces the variation of identical pedestrians in the sample and facilitates the learning of similarity features. Both types of methods mentioned above process each video frame independently. The features extracted by LSTM are generally affected by the length of the video sequence. The RNN only establishes temporal associations on high-level features and thus cannot capture the temporal cues of local details of the image [167]. Therefore, there is still a need to explore a more efficient method for extracting spatial-temporal features.

3.4.4. Spatial-temporal attention

The attention mechanism can selectively focus on the useful local information and has good performance in solving the problems of camera view switching, lighting changes and occlusion in person Re-ID tasks. Recently, several researchers have used attentional mechanisms to solve video-based person Re-ID tasks in both temporal and spatial dimensions. To solve the problem of unalignment and occlusion caused by changes in pedestrian body pose and camera view in video sequences, Li et al. [170] proposed a spatial-temporal attention model, the core idea of which is to use multiple spatial attention to extract features of key body parts and use temporal attention to compute the combined feature representations extracted by each spatial attention model. This method can better mine the potential distinguishing feature representations in video sequences. Similarly, Fu et al. [58] proposed a spatial-temporal attention framework that can fully utilize the distinguishing features of each pedestrian in both temporal and spatial dimensions through video frame selection, local feature mining, and feature fusion. The approach can well address challenges such as pedestrian pose variation and partial occlusion. Xu et al. [166] proposed a joint temporal and spatial attention pooling network to learn the feature representations of video sequences through the interdependence between video sequences.

3.4.5. Graph convolutional networks

In recent years, graph convolutional networks (GCNs) have been widely used for person Re-ID tasks due to their powerful automatic relational modeling capabilities [175], and a large number of variant networks have emerged [60,62–65]. Yang et al. [60] proposed a unified spatial-temporal graph convolutional neural network that modelled video sequences in three dimensions: temporal, spatial and appearance, and to mine more discriminative and robust information. Wu et al. [62]

proposed an adaptive graph representation learning scheme for video person Re-ID using pose alignment connections and feature similarity connections to construct adaptive structure-aware adjacency graphs.

Yan et al. [63] proposed a framework for pedestrian retrieval based on contextual graphical convolutional networks. Since image appearance features are not sufficient to distinguish different people, the authors use contextual information to extend instance-level features to improve the discriminative power of the features and the robustness of person retrieval. Shen et al. [64] proposed a similarity-guided graph neural network that represents pairwise relationships between probe-gallery image pairs (nodes) by creating a graph and using this relationship to enhance the learning of discriminative features. This updated probe-gallery image is used to predict the relational features for accurate similarity estimation.

The last group of Table 7 shows the experimental results of the GCN-based sequence feature learning methods on MARS, DukeMTMC-VideoReID iLIDS-VID and PRID-2011 datasets. From the results in the above table, the experimental performance of the GCN-based methods is significantly better than the other sequence feature learning methods. In particular, CTL [174] achieved Rank-1 accuracy of 91.4% and mAP of 86.7% on MARS. CTL utilized a CNN backbone and a key-points estimator to extract semantic local features from the human body at multiple granularities as graph nodes. CTL effectively mined comprehensive cues complementary to appearance information to enhance the representation capability.

3.4.6. Comparison and discussion

Table 7 shows the experimental results of sequence feature learning-based methods on MARS, DukeMTMC-VideoReID, iLIDS-VID and PRID-2011 datasets. Some researchers used GCN to model the Spatio-temporal relationships of video sequences and achieved good results. Compared with optical flow, 3DCNN, and RNN or LSTM methods, spatial-temporal attention-based methods can obtain better experimental performance. AdaptiveGraph [62], MGH [61] and STGCN [60] were able to obtain high experimental results on the above datasets, with some improvement in accuracy compared to the previous types of methods.

The core idea of the sequence feature learning-based methods is to fuse more spatial-temporal information from multiple dimensions to mitigate the effects of a range of problems such as occlusion, illumination, and viewpoint changes in the person Re-ID tasks. Although Optical flow can provide contextual information of video sequence frames, it

only represents the local dynamics of adjacent images, which may introduce noise due to spatial misalignment. The process of computing optical flow is time-consuming. 3DCNN can capture both temporal and spatial feature information in video sequences. Although 3DCNN can achieve better performance, these methods are computationally time-consuming and hard to optimize. RNN or LSTM can extract temporal features of video sequences and has been popular among researchers for some time. In the task of person Re-ID, the ability of RNN or LSTM have limited features for temporal information extraction and suffer from difficulties in model training due to the complex network structure [176]. Although the introduction of temporal attention and spatial attention can alleviate the problem of switching between different camera views in a row, lighting changes and occlusion, the accuracy of person Re-ID is affected because of the temporal relationship between body parts in different frames is not fully considered [60]. Most of the current person Re-ID studies are image-based, and a large number of methods and datasets closer to the real world have emerged. Compared to image-based methods, sequence feature learning-based methods still hold great research promise.

4. Conclusion and future directions

This paper presents a comprehensive survey with in-depth discussion for deep learning-based person Re-ID methods in recent years. Firstly, we summarize the main contributions of several recently published person Re-ID surveys and discuss the common datasets used for person Re-ID benchmarks. Secondly, we comprehensively review the current deep learning-based methods, these methods are classified into four major categories according to metric learning and representation learning, including deep metric learning, local feature learning, generative adversarial learning and sequence feature learning. We subdivide the above four categories according to their methodologies and motivations, analyzing and discussing the advantages and limitations of each subcategory of the method. This classification is more suitable for researchers to explore these methods from their practical needs.

Although existing deep learning-based methods have achieved good results in person Re-ID tasks, they still face many challenges. Most of the datasets currently applied for person Re-ID training are processed visible images or videos, but real-world data often exhibit a combination of multiple modalities. Although semi-supervised and unsupervised methods can alleviate the problem of high labeling costs, they still do not perform as well as supervised methods. There are domain

Table 7

Comparison of experimental results of video sequence feature learning-based methods. * represents the use of multiple deep learning methods. The bold and underlined numbers represent the top two results.

Categories	Methods	MARS		DukeMTMC-VideoReID		iLIDS-VID		PRID-2011		Reference
		mAP	R-1	mAP	R-1	R-1	R-5	R-1	R-5	
Optical flow	RCN [165]*	–	–	–	–	58.0	84.0	70.0	90.0	CVPR'16
	TSSCNN [56]	–	–	–	–	60.0	86.0	78.0	94.0	ICCV'17
	ASTPN [166]	–	44.0	–	–	62.0	86.0	70.0	90.0	ICCV'17
	AMOC [163]	52.9	68.3	–	–	68.7	94.3	83.7	98.3	TCSVT'18
	CSACSE [164]*	69.4	81.2	–	–	79.8	91.8	81.2	92.1	CVPR'18
3DCNN	CSACSE + OF [164]*	76.1	86.3	–	–	85.4	96.7	93.0	99.3	CVPR'18
	3DCNN + NLA [57]	77.0	84.3	–	–	81.3	–	91.2	–	ACCV'18
RNN or LSTM	M3D [167]	74.1	84.4	–	–	74.0	94.3	94.4	100.0	AAAI'19
	RFA [169]	–	–	–	–	49.3	76.8	58.2	85.8	ECCV'16
	SFT [168]	50.7	70.6	–	–	55.2	86.5	79.4	94.4	CVPR'17
Spatial-temporal attention	DRSA [170]	65.8	82.3	–	–	80.2	–	93.2	–	CVPR'18
	GLTR [172]	78.5	87.0	93.7	96.3	86.0	98.0	95.5	100.0	ICCV'19
	VRSTC [59]	82.3	88.5	93.5	95.0	83.4	95.5	–	–	CVPR'19
	STA [58]	80.8	86.3	94.9	96.2	–	–	–	–	AAAI'19
	MG-RAFA [171]	85.6	88.8	–	–	88.6	98.0	95.9	99.7	CVPR'20
GCN	BiCnet-TKS [173]	86.0	90.2	96.1	96.3	75.1	84.6	–	–	CVPR'21
	AdaptiveGraph [62]	81.9	89.5	95.4	97.0	84.5	96.7	94.6	99.1	TIP'20
	MGH [61]	85.8	90.0	–	–	85.6	97.1	94.8	99.3	CVPR'20
	STGCN [60]	83.7	89.9	95.7	97.3	–	–	–	–	CVPR'20
	CTL [174]	86.7	91.4	–	–	89.7	–	–	–	CVPR'21

differences in pedestrian images captured by different cameras, and models trained on one dataset can experience severe performance degradation when tested on another dataset. Because people may change their clothes or different people may wear very similar clothes, the appearance features of pedestrians will become unreliable for person Re-ID. In addition, how to improve the speed and accuracy of model retrieval is critical for real-world model deployment. The increase of privacy scenarios fundamentally limits the traditional centralized person Re-ID methods. The detection and re-identification modules of most person Re-ID systems are separated from each other, making it difficult to expand to real-world applications.

In summary, there are still many challenges to be explored and researched in deep learning-based person Re-ID methods. The following subsections present potential solutions to address the above existing challenges, as well as prospects of future research directions.

- (1) **Cross-modal person Re-ID.** Most existing Re-ID methods evaluate their performance on publicly available datasets, which are obtained based on image or video processing. However, the acquisition of real-world data is diverse and the data may appear as a combination of different modalities (visible, infrared, depth map and text descriptions, etc.). For example, in the absence of sufficient visibility information like images or videos, text descriptions can provide unique attributes aiding information for person Re-ID. Several research works [161,177] learned discriminative cross-modal visual-textual features for better similarity evaluation in description-based person Re-ID. Because it was difficult for visible-light cameras to capture valid appearance information in dark environments, some researchers [178–182] used thermal infrared images to learn rich visual representations for cross-modality matching. Existing works mainly focus on alleviating the modality discrepancy by aligning the distributions of features from different modalities. Meanwhile, how to combine various modal complementary information is also worth studying in the future.
- (2) **High-performance semi-supervised and unsupervised person Re-ID.** Because it was expensive to annotate person images across multiple cameras, some researchers [183–188] focused on semi-supervised and unsupervised methods for person Re-ID. These methods aimed to learn discriminative features from unlabeled or minimally labeled images of people. Compared to supervised learning, semi-supervised and unsupervised methods alleviated the need for expensive data annotation and showed great potential to facilitate person Re-ID to practical applications. Some semi-supervised person Re-ID methods [189,190] made use of image clustering or tracklet clustering in the target domain to adapt the model to the new domain. Some unsupervised person Re-ID methods [183–185] used soft labels or multi-labels to learn discriminative embedding features. Although lacking realistic label learning discriminative features, the performance of person Re-ID methods in semi-supervised and unsupervised scenarios was still inferior to that of supervised methods, they still maintained significant research value and significance in improving the generalization ability of the model [2]. In future research, better clustering or label assignment strategies should be considered to improve the performance of person Re-ID.
- (3) **Domain adaptation person Re-ID.** The background, resolution, and illumination environment of different cameras in the real world vary greatly, which interfered with the learning of distinguishing features of pedestrians and affected the performance of person Re-ID. Some researchers [45,47–49] transferred images with identity labels from the source to the target domain to learn discriminative models, but they largely ignored the unlabeled samples and the substantial sample distributions in target domains. Some researchers [155,191–194] used clustering or graph matching methods to predict the pseudo-labels in the

target domain for discriminative model learning, but they still faced the challenge of accurately predicting hard samples labels. Domain adaption is crucial for person Re-ID models learned in unknown domains. Therefore, it remains one of the important research directions for the future.

- (4) **Person Re-ID in the 3D space.** In the real world, the spatial location of cameras is uncertain and a new camera may be temporarily inserted into an existing camera network. Considering people may change their clothes or different people may wear very similar clothes, pedestrian appearance features will become unreliable for Re-ID [67]. The 3D structure does not rely on the appearance information of 2D images can effectively alleviate this limitation. However, the acquisition of 3D point cloud data of pedestrians requires additional auxiliary models. Some researchers [66,67] extracted 3D shape embeddings directly from 2D images, obtained more robust structural and appearance information by aligning 2D and 3D local features, or planed 3D models back to 2D images for representation learning in 2D space for data augmentation purposes. Although the above studies achieved good experimental results, the 2D data space inherently limited the model to understand the 3D geometric information of people. Therefore, further exploration of person Re-ID methods in 3D space is still an important research direction in the future.
- (5) **Fast person Re-ID.** Most current methods to person Re-ID focus mainly on prior knowledge or designing complex network architectures to learn robust identity invariant feature representations. These methods use complex network models to extract high-dimensional features to improve model performance. However, the above methods use Euclidean distance to calculate the similarity of features and obtain the rank list by fast sorting, which will increase with the retrieval time as the size of the gallery library increases. This retrieval method will be very time-consuming, making the model unsuitable for real-world applications. Therefore, some researchers [195–197] considered introducing hashing to improve the retrieval speed. Faster person Re-ID retrieval from coarse-to-fine (CtF) [196] can be achieved by supplementing long and short hash codes to get faster and better accuracy. Zhao et al. [197] proposed saliency-guided iterative asymmetric mutual hashing (SIAMH) to achieve high-quality hash code generation and fast feature extraction. However, how to design a specific retrieval strategy to reduce the information redundancy among models and improve retrieval speed and accuracy still needs further research.
- (6) **Decentralised learning person Re-ID.** Most of the existing person Re-ID methods use a centralized learning paradigm, which requires collecting all training data from different camera views or domains for centralized training. Although these supervised or unsupervised methods have made significant progress, centralized person Re-ID learning ignores images of people that contain large amounts of personal and private information that may not be allowed to be shared into a central data set. As privacy scenarios increase, it can fundamentally limit the centralized learning person Re-ID methods in the real world. Several recent works [198–201] attempted to address the above problem through decentralised learning. These methods either built a globally generalised model server through federated learning, which did not require access to local training data and shared of cross-domain data, or selectively performed knowledge aggregation to optimize the trade-off between model personalisation and generalization in decentralised person Re-ID. In future work, how to ensure understanding cross-domain data heterogeneity while learning a global generalised model remains challenging.
- (7) **End-to-end person Re-ID system.** As is shown in Fig. 2 in the introduction of this paper, person detection and re-identification in most current person Re-ID systems are two independent

modules. The person Re-ID task uses the correct pedestrian already detected by default, but some practical open-world applications require end-to-end person search from the raw images or videos [202]. The two-stage end-to-end person re-identification framework was one of the most common Re-ID systems that systematically evaluated the advantages and limitations of combining different detectors and Re-ID models [112]. Munjal et al. [203] proposed a query-guided end-to-end person search network (QEEPS) to join person detection and re-identification. In addition, the end-to-end person Re-ID was also widely used in multi-target multi-camera tracking (MTMC tracking) [110,204,205]. Person Re-ID algorithms rely not only on accurate person detection algorithms but also on detected unlabeled pedestrians, which remains a current challenge. Therefore, how to effectively combine person detection and re-identification to design an end-to-end person Re-ID system is also a direction that researchers need to pay attention to in the future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors wish to thank Jizhuo Li, Xiao Pang, Jiamin Zhu, Fuqiu Chen, Yi Zhou and Cheng Zhang. This work was supported by the National Key Research and Development Project of China (No. JG2018190), and in part by the National Natural Science Foundation of China (No. 61872256).

References

- [1] X. Wang, Intelligent multi-camera video surveillance: a review, *Pattern Recogn. Lett.* 34 (2013) 3–19.
- [2] H. Luo, W. Jiang, X. Fan, S. Zhang, A survey on deep learning based person re-identification, *Acta Automat. Sin.* 45 (2019) 2032–2049.
- [3] L. Zheng, Y. Yang, A.G. Hauptmann, Person Re-identification: Past, Present and Future, <https://arxiv.org/pdf/1610.02984.pdf> 2016.
- [4] J. Redmon, A. Farhadi, Yolov3: An Incremental Improvement, <https://arxiv.org/pdf/1804.02767.pdf> 2018 (arXiv preprint arXiv:1804.02767).
- [5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. Berg, Ssd: single shot multibox detector, *Proceedings of the European Conference on Computer Vision* 2016, pp. 21–37.
- [6] R. Girshick, Fast r-cnn, *Proceedings of the IEEE International Conference on Computer Vision* 2015, pp. 1440–1448.
- [7] L. Qi, P. Yu, Y. Gao, Research on weak-supervised person re-identification, *J. Softw.* 9 (2020) 2883–2902.
- [8] Z. Zhong, L. Zheng, S. Cao, D. Li, Re-ranking person re-identification with k-reciprocal encoding, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017, pp. 3652–3661.
- [9] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, *Proceedings of the IEEE International Conference on Computer Vision* 2013, pp. 2528–2535.
- [10] N. Martinel, C. Micheloni, G. Foresti, Saliency weighted features for person re-identification, *Proceedings of the European Conference on Computer Vision* 2015, pp. 191–208.
- [11] L. An, X. Chen, S. Liu, Y. Lei, S. Yang, Integrating appearance features and soft biometrics for person re-identification, *Multimed. Tools Appl.* 76 (2017) 12117–12131.
- [12] H. Hu, W. Fang, G. Zeng, Z. Hu, B. Li, A person re-identification algorithm based on pyramid color topology feature, *Multimed. Tools Appl.* 76 (2017) 26633–26646.
- [13] M. Dikmen, E. Akbas, T. Huang, N. Ahuja, Pedestrian recognition with a learned metric, *Proceedings of the Asian Conference on Computer Vision* 2011, pp. 501–512.
- [14] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: deep filter pairing neural network for person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2014, pp. 152–159.
- [15] Y. Chen, W. Zheng, J. Lai, Mirror representation for modeling view-specific transform in person re-identification, *Proceedings of the International Joint Conference on Artificial Intelligence* 2015, pp. 3402–3408.
- [16] X. Wang, W. Zheng, X. Li, J. Zhang, Cross-scenario transfer person reidentification, *IEEE Transactions on Circuits and Systems for Video Technology*, 26, 2016, pp. 1447–1460.
- [17] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, *Proceedings of the International Conference on Pattern Recognition* 2014, pp. 34–39.
- [18] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: deep filter pairing neural network for person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2014, pp. 152–159.
- [19] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, pp. 5177–5186.
- [20] M. Zheng, S. Karanam, Z. Wu, R. Radke, Re-identification with consistent attentive siamese networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 5735–5744.
- [21] R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, *Proceedings of the European Conference on Computer Vision* 2016, pp. 791–808.
- [22] A. Hermans, L. Beyer, B. Leibe, In Defense of the Triplet Loss for Person Re-identification, <https://arxiv.org/pdf/1703.07737.pdf> 2017 (arXiv preprint arXiv: 1703.07737).
- [23] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017, pp. 403–412.
- [24] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline), *Proceedings of the European Conference on Computer Vision* 2018, pp. 480–496.
- [25] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, J. Sun, Perceive where to focus: learning visibility-aware part-level features for partial person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 393–402.
- [26] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Wang, Abd-net: attentive but diverse person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2019, pp. 8350–8360.
- [27] J. Miao, Y. Wu, P. Liu, Y. Ding, Y. Yang, Pose-guided feature alignment for occluded person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2019, pp. 542–551.
- [28] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017, pp. 1077–1085.
- [29] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, X. Wang, Hydraplus-net: attentive deep features for pedestrian analysis, *Proceedings of the IEEE International Conference on Computer Vision* 2017, pp. 350–359.
- [30] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, pp. 2285–2294.
- [31] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2017, pp. 3239–3248.
- [32] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, pp. 7132–7141.
- [33] B. Chen, W. Deng, J. Hu, Mixed high-order attention network for person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2019, pp. 371–381.
- [34] C.P. Tay, S. Roy, K.H. Yap, Aanet: attribute attention network for person re-identifications, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 7127–7136.
- [35] C.W. Ngo, Z. Pan, X. Wei, X. Wu, H.K. Tan, W. Zhao, Motion Driven Approaches to Shot Boundary Detection, Low-level Feature Extraction and BBC Rushes Characterization at TRECVID 2005, *TRECVID Workshop Participants Notebook Papers*, 2005.
- [36] C.W. Ngo, Y.G. Jiang, X.Y. Wei, W. Zhao, F. Wang, X. Wu, H.K. Tan, Beyond Semantic Search: What you Observe May Not be What You Think, *TRECVID Workshop Participants Notebook Papers*, 2008.
- [37] X.Y. Wei, Z.Q. Yang, Mining in-class social networks for large-scale pedagogical analysis, *Proceedings of the ACM international conference on Multimedia* 2012, pp. 639–648.
- [38] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, *Proceedings of the ACM International Conference on Multimedia* 2018, pp. 274–282.
- [39] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, Glad: global-local-alignment descriptor for pedestrian retrieval, *Proceedings of the ACM International Conference on Multimedia* 2017, pp. 420–428.
- [40] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2017, pp. 3980–3989.
- [41] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, Y. Yang, Salience-guided cascaded suppression network for person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2020, pp. 3297–3307.
- [42] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2020, pp. 3183–3192.
- [43] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, R. Ji, Pyramidal person re-identification via multi-loss dynamic training, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 8506–8514.
- [44] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, Q. Tian, Deep representation learning with part loss for person re-identification, *IEEE Trans. Image Process.* 28 (2019) 2860–2871.

- [45] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 79–88.
- [46] Y. Huang, Z.J. Zha, X. Fu, R. Hong, L. Li, Real-world person re-identification via degradation invariance learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2020, pp. 14084–14094.
- [47] W. Deng, Z. Liang, G. Kang, Y. Yi, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 994–1003.
- [48] J. Liu, Z. Zha, D. Chen, R. Hong, M. Wang, Adaptive transfer network for cross-domain person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 7202–7211.
- [49] Y. Chen, X. Zhu, S. Gong, Instance-guided context rendering for cross-domain person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2020, pp. 232–242.
- [50] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero and homogeneously, Proceedings of the European Conference on Computer Vision 2018, pp. 172–188.
- [51] Z. Zhong, Z. Liang, Z. Zheng, S. Li, Y. Yi, Camera style adaptation for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 5157–5166.
- [52] Z. Zheng, Z. Liang, Y. Yi, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 618–626.
- [53] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, J. Zhang, Multi-pseudo regularized label for generated data in person re-identification, IEEE Trans. Image Process. 28 (2019) 1391–1403.
- [54] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 2138–2147.
- [55] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, H. Li, Fd-gan: Pose-guided Feature Distilling Gan for Robust Person Re-identification, <https://arxiv.org/pdf/1810.02936.pdf> 2018 (arXiv preprint arXiv:1810.02936).
- [56] D. Chung, K. Tahboub, E. Delp, A two stream siamese convolutional neural network for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 1983–1991.
- [57] X. Liao, L. He, Z. Yang, C. Zhang, Video-based person re-identification via 3d convolutional networks and non-local attention, Proceedings of the Asian Conference on Computer Vision 2019, pp. 620–634.
- [58] Y. Fu, X. Wang, Y. Wei, T. Huang, Sta: spatial-temporal attention for large-scale video-based person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8287–8294.
- [59] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Vrsc: occlusion-free video person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 7183–7192.
- [60] J. Yang, W. Zheng, Q. Yang, Y. Chen, Q. Tian, Spatial-temporal graph convolutional network for video-based person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2020, pp. 3289–3299.
- [61] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, L. Shao, Learning multi-granular hypergraphs for video-based person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2020, pp. 2899–2908.
- [62] Y. Wu, O. Bourahla, X. Li, F. Wu, Q. Tian, X. Zhou, Adaptive graph representation learning for video person re-identification, IEEE Trans. Image Process. 29 (2020) 8821–8830.
- [63] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, X. Yang, Learning context graph for person search, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 2158–2167.
- [64] Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, Proceedings of the European Conference on Computer Vision 2018, pp. 486–504.
- [65] Z. Bai, Z. Wang, J. Wang, D. Hu, E. Ding, Unsupervised multi-source domain adaptation for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2021, pp. 12914–12923.
- [66] Z. Zheng, N. Zheng, Y. Yang, Parameter-efficient Person Re-identification in the 3d space, <https://arxiv.org/pdf/2006.04569.pdf> 2020 (arXiv preprint arXiv:2006.04569).
- [67] J. Chen, X. Jiang, F. Wang, J. Zhang, F. Zheng, X. Sun, W.S. Zheng, Learning 3d shape feature for texture-insensitive person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2021, pp. 8146–8155.
- [68] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-identification, Image Vis. Comput. 32 (2014) 270–286.
- [69] H. Chahar, N. Nain, A study on deep convolutional neural network based approaches for person re-identification, Pattern Recognition and Machine Intelligence 2017, pp. 543–548.
- [70] K. Wang, H. Wang, M. Liu, X. Xing, T. Han, Survey on person re-identification based on deep learning, CAAI Transactions on Intelligence Technology, 3, 2018, pp. 219–227.
- [71] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. Radke, A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 523–536.
- [72] Z. Wang, Z. Wang, Y. Zheng, Y. Wu, W. Zeng, S. Satoh, Beyond Intra-modality: A Survey of Heterogeneous Person Re-identification, <https://arxiv.org/pdf/1905.10048.pdf> 2020 (arXiv preprint arXiv:1905.10048).
- [73] D. Wu, S. Zheng, X. Zhang, C. Yuan, F. Cheng, Y. Zhao, Y. Lin, Z. Zhao, Y. Jiang, D. Huang, Deep learning-based methods for person re-identification: a comprehensive review, Neurocomputing 337 (2019) 354–371.
- [74] N. Mathur, S. Mathur, D. Mathur, P. Dadhee, A brief survey of deep learning techniques for person re-identification, Proceedings of International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things 2020, pp. 129–138.
- [75] Q. Leng, M. Ye, Q. Tian, A survey of open-world person re-identification, IEEE Transactions on Circuits and Systems for Video Technology, 30, 2020, pp. 1092–1108.
- [76] K. Islam, Person search: New paradigm of person re-identification: a survey and outlook of recent works, Image Vis. Comput. 101 (2020), 103970.
- [77] B. Lavi, I. Ullah, M. Fatan, A. Rocha, Survey on Reliable Deep Learning-based Person Re-identification Models: Are We There Yet? <https://arxiv.org/pdf/2005.00355.pdf> 2020 (arXiv preprint arXiv:2005.00355).
- [78] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
- [79] C. Yang, F. Qi, H. Jia, Survey on unsupervised techniques for person re-identification, International Conference on Computing and Data Science 2021, pp. 161–164.
- [80] E. Yaghoubi, A. Kumar, H. Proen  a, Sss-pr: a short survey of surveys in person re-identification, Pattern Recogn. Lett. 143 (2021) 50–57.
- [81] Y. Wang, S. Yang, S. Liu, Z. Zhang, Cross-domain person re-identification: a review, Artif. Intell. China (2021) 153–160.
- [82] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, C.C. Loy, Domain Generalization: A Survey, <https://arxiv.org/pdf/2103.02503.pdf> 2021.
- [83] N.K.S. Behera, P.K. Sa, S. Bakshi, Person re-identification for smart cities: State-of-the-art and the path ahead, Pattern Recogn. Lett. 138 (2020) 282–289.
- [84] W. Wu, D. Tao, H. Li, Z. Yang, J. Cheng, Deep features for person re-identification on metric learning, Pattern Recogn. 110 (2021), 107424.
- [85] N.K.S. Behera, T.K. Behera, M. Nappi, S. Bakshi, P.K. Sa, Futuristic person re-identification over internet of biometrics things (iobt): technical potential versus practical reality, Pattern Recogn. Lett. 151 (2021) 163–171.
- [86] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.
- [87] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, IEEE International Workshop on Performance Evaluation for Tracking and Surveillance 2007, pp. 1–7.
- [88] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, Proceedings of the Asian conference on Computer Vision 2012, pp. 31–44.
- [89] W. Li, X. Wang, Locally aligned feature transforms across views, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2013, pp. 3594–3601.
- [90] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1627–1645.
- [91] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 1116–1124.
- [92] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 3774–3782.
- [93] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target multi-camera tracking, Proceedings of the European Conference on Computer Vision 2016, pp. 17–35.
- [94] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Proces. Syst. 39 (2015) 1137–1149.
- [95] P. Dollar, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1532–1545.
- [96] M. Hirzer, C. Beleznai, P. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, Image Analysis (2011) 91–102.
- [97] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, Proceedings of the European Conference on Computer Vision 2014, pp. 688–703.
- [98] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: a video benchmark for large-scale person re-identification, Proceedings of the European Conference on Computer Vision 2016, pp. 868–884.
- [99] G. Song, B. Leng, Y. Liu, C. Hetang, S. Cai, Region-based quality estimation network for large-scale person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence 2018, pp. 7347–7354.
- [100] A. Dehghan, S. Modiri Assari, M. Shah, Gmmcp tracker: globally optimal generalized maximum multi clique problem for multiple object tracking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 4091–4099.
- [101] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4004–4012.
- [102] Y. Duan, J. Lu, J. Feng, J. Zhou, Deep localized metric learning, IEEE Transactions on Circuits and Systems for Video Technology, 28, 2018, pp. 2644–2656.
- [103] H. Huang, D. Li, Z. Zhang, X. Chen, K. Huang, Adversarially occluded samples for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 5098–5107.
- [104] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang, Invariance matters: exemplar memory for domain adaptive person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 598–607.

- [105] C. Luo, Y. Chen, N. Wang, Z. Zhang, Spectral feature transformation for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 4976–4985.
- [106] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person re-identification, ACM Trans. Multimed. Comput. Commun. Appl. 14 (2017) 1–20.
- [107] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian, T. Xiang, Deep transfer learning for person re-identification, Proceedings of the IEEE International Conference on Multimedia Big Data 2018, pp. 1–5.
- [108] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, Proceedings of the European Conference on Computer Vision 2016, pp. 135–153.
- [109] Y. Wang, Z. Chen, F. Wu, G. Wang, Person re-identification with cascaded pairwise convolutions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 1470–1478.
- [110] E. Ristani, C. Tomasi, Features for multi-target multi-camera tracking and re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 6036–6046.
- [111] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 1179–1188.
- [112] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 1367–1376.
- [113] H. Li, G. Wu, W.S. Zheng, Combined depth space based architecture search for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2021, pp. 6729–6738.
- [114] J. Guo, Y. Yuan, L. Huang, C. Zhang, J. Yao, K. Han, Beyond human parts: dual part-aligned representations for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 3642–3651.
- [115] G. Chen, C. Lin, L. Ren, J. Lu, J. Zhou, Self-critical attention learning for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 9637–9646.
- [116] Z. Liu, J. Wang, S. Gong, H. Lu, D. Tao, Deep reinforcement active learning for human-in-the-loop person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 6122–6131.
- [117] J. Song, Y. Yang, Y. Song, T. Xiang, T. Hospedales, Generalizable person re-identification by domain-invariant mapping network, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 719–728.
- [118] S. Zhou, F. Wang, Z. Huang, J. Wang, Discriminative feature learning with consistent attention regularization for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 8040–8049.
- [119] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S.Z. Li, Embedding deep metric for person re-identification: a study against large variations, Proceedings of the European Conference on Computer Vision 2016, pp. 732–748.
- [120] A. Mishchuk, D. Mishkin, F. Radenovic, J. Matas, Working hard to know your neighbor's margins: local descriptor learning loss, Proceedings of the International Conference on Neural Information Processing Systems 2017, pp. 4829–4840.
- [121] C. Yan, G. Pang, X. Bai, C. Liu, N. Xin, L. Gu, J. Zhou, Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss, IEEE Trans. Multimed. (2021).
- [122] L. Wu, C. Shen, A. Hengel, Personnet: Person Re-identification With Deep Convolutional Neural Networks, <https://arxiv.org/pdf/1601.07255.pdf> 2016 (arXiv preprint arXiv:1601.07255).
- [123] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 1288–1296.
- [124] X. Qian, Y. Fu, Y. Jiang, T. Xiang, X. Xue, Multi-scale deep learning architectures for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 5399–5408.
- [125] J. Liu, Z. Zha, H. Xie, Z. Xiong, Y. Zhang, Ca3net: contextual-attentional attribute-appearance network for person re-identification, Proceedings of the ACM international conference on Multimedia 2018, pp. 737–745.
- [126] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8295–8302.
- [127] J. Liu, Z. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, T. Mei, Multi-scale triplet cnn for person re-identification, Proceedings of the ACM International Conference on Multimedia 2016, pp. 192–196.
- [128] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, Proceedings of the IEEE International Conference on Computer Vision Workshops 2017, pp. 2590–2600.
- [129] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 3702–3712.
- [130] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 384–393.
- [131] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, S. Zhang, Towards rich feature discovery with class activation maps augmentation for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 1389–1398.
- [132] Z. Zhang, H. Zhang, S. Liu, Person re-identification using heterogeneous local graph attention networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2021, pp. 12136–12145.
- [133] X. Ning, K. Gong, W. Li, L. Zhang, Jwsaa: joint weak saliency and attention aware for person re-identification, Neurocomputing 453 (2021) 801–811.
- [134] M.M. Kalayeh, E. Basaran, M. Gökmén, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 1062–1071.
- [135] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Learning generalisable omni-scale representations for person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
- [136] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, S. Tian, Feature refinement and filter network for person re-identification, IEEE Transactions on Circuits and Systems for Video Technology, 31, 2021, pp. 3391–3402.
- [137] Y. Ming, Z. Yang, X. Wei, J. Yan, X. Wang, F. Wang, M. Zhu, Global-local Dynamic Feature Alignment Network for Person Re-identification, <https://arxiv.org/pdf/2109.05759.pdf> 2021 (arXiv preprint arXiv:2109.05759).
- [138] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, Diverse part discovery: occluded person re-identification with part-aware transformer, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2021, pp. 2898–2907.
- [139] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Proceedings of the International Conference on Neural Information Processing Systems, 2, 2014, pp. 2672–2680.
- [140] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, J. Hu, Pose transferable person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 4099–4108.
- [141] Y.C. Chen, Y.J. Li, Y.C.F. Du, X. Wang, Learning resolution-invariant deep representations for person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8215–8222.
- [142] J. Zhu, T. Park, P. Isola, A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 2223–2232.
- [143] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, J. Choo, Stargan: unified generative adversarial networks for multi-domain image-to-image translation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 8789–8797.
- [144] W. Liang, G. Wang, J. Lai, J. Zhu, M2m-gan: Many-to-many Generative Adversarial Transfer Learning for Person Re-identification, <https://arxiv.org/pdf/1811.03768.pdf> 2018 (arXiv preprint arXiv:1811.03768).
- [145] C. Liu, X. Chang, Y.D. Shen, Unity style transfer for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2020, pp. 6887–6896.
- [146] Z. Wang, Z. Wang, Y. Zheng, Y.Y. Chuang, S.I. Satoh, Learning to reduce dual-level discrepancy for infra-red-visible person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 618–626.
- [147] S. Bak, P. Carr, J.F. Lalonde, Domain adaptation through synthesis for unsupervised person re-identification, Proceedings of the European Conference on Computer Vision 2018, pp. 189–205.
- [148] C. Dai, H. Wang, T. Ni, S. Chen, Person re-identification based on deep convolutional generative adversarial network and expanded neighbor reranking, J. Comput. Res. Dev. 56 (2019) 1632–1641.
- [149] W.X. Yang, Y. Yan, S. Chen, X.K. Zhang, H.Z. Wang, Multi-scale generative adversarial network for person re-identification under occlusion, J. Comput. Res. Dev. 31 (2020) 1943–1958.
- [150] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, X. Xue, Pose-normalized image generation for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 650–667.
- [151] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, X. Bai, Progressive pose attention transfer for person image generation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 2347–2356.
- [152] L. Ma, Q. Sun, S. Georgoulis, L.V. Gool, B. Schiele, M. Fritz, Disentangled person image generation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 99–108.
- [153] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, F. Bremond, Joint generative and contrastive learning for unsupervised person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2021, pp. 2004–2013.
- [154] Y.J. Li, Y.C. Chen, Y.Y. Lin, X. Du, Y.C.F. Wang, Recover and identify: a generative dual model for cross-resolution person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 8090–8099.
- [155] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: clustering and fine-tuning, ACM Trans. Multimed. Comput. Commun. Appl. 14 (2018) 1–18.
- [156] A. Torralba, A.A. Efros, Unbiased look at dataset bias, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2011, pp. 1521–1528.
- [157] J. Liu, Z. Zha, R. Hong, M. Wang, Y. Zhang, Deep adversarial graph attention convolution network for text-based person search, Proceedings of the ACM International Conference on Multimedia 2019, pp. 665–673.
- [158] S. Mao, S. Zhang, M. Yang, Resolution-invariant Person Re-identification, <https://arxiv.org/pdf/1906.09748.pdf> 2019 (arXiv preprint arXiv:1906.09748).
- [159] S. Liao, H. Yang, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 2197–2206.
- [160] P. Peng, X. Tao, Y. Wang, M. Pontil, Y. Tian, Unsupervised cross-dataset transfer learning for person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 1306–1315.
- [161] Z. Zha, J. Liu, D. Chen, F. Wu, Adversarial attribute-text embedding for person search with natural language query, IEEE Trans. Multimed. 22 (2020) 1836–1846.
- [162] Y. Zou, X. Yang, Z. Yu, B. Kumar, J. Kautz, Joint disentangling and adaptation for cross-domain person re-identification, Proceedings of the European Conference on Computer Vision 2020, pp. 87–104.

- [163] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, J. Feng, Video-based person re-identification with accumulative motion context, *IEEE Trans. Circuits Syst. Video Technol.* 28 (2018) 2788–2802.
- [164] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, pp. 1169–1178.
- [165] N. McLaughlin, J. del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2016, pp. 1325–1334.
- [166] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, P. Zhou, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2017, pp. 4733–4742.
- [167] J. Li, S. Zhang, T. Huang, Multi-scale 3d convolution network for video based person re-identification, *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 2019, pp. 8618–8625.
- [168] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017, pp. 4747–4756.
- [169] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, X. Yang, Person re-identification via recurrent feature aggregation, *Proceedings of the European Conference on Computer Vision* 2016, pp. 701–716.
- [170] S. Li, S. Bak, P. Carr, X. Wang, Diversity regularized spatiotemporal attention for video-based person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018, pp. 369–378.
- [171] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2020, pp. 10407–10416.
- [172] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2019, pp. 3958–3967.
- [173] R. Hou, H. Chang, B. Ma, R. Huang, S. Shan, Bicnet-tks: learning efficient spatial-temporal representation for video person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2021, pp. 2014–2023.
- [174] J. Liu, Z.J. Zha, W. Wu, K. Zheng, Q. Sun, Spatial-temporal correlation and topology learning for person re-identification in videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2021, pp. 4370–4379.
- [175] T. Kipf, M. Welling, Semi-supervised Classification With Graph Convolutional Networks, <https://arxiv.org/pdf/1609.02907.pdf> 2017 (arXiv preprint arXiv: 1609.02907).
- [176] J. Gao, R. Nevatia, Revisiting Temporal Modeling for Video-based Person Reid, <https://arxiv.org/pdf/1805.02104.pdf> 2018 (arXiv preprint arXiv:1805.02104).
- [177] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, *IEEE Trans. Image Process.* 29 (2020) 5542–5556.
- [178] S. Choi, S. Lee, T. Kim, Y. Kim, C. Kim, Hi-cmd: hierarchical cross-modality disentanglement for visible-infrared person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2020, pp. 10257–10266.
- [179] M. Ye, J. Shen, D.J. Crandall, L. Shao, J. Luo, Dynamic dual-attentive aggregation learning for visible-infrared person re-identification, *Proceedings of the European Conference on Computer Vision* 2020, pp. 229–247.
- [180] G.A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, Z.G. Hou, Cross-modality paired-images generation for rgb-infrared person re-identification, *Proceedings of the AAAI Conference on Artificial Intelligence* 2020, pp. 12144–12151.
- [181] D. Li, X. Wei, X. Hong, Y. Gong, Infrared-visible cross-modal person re-identification with an x modality, *Proceedings of the AAAI Conference on Artificial Intelligence* 2020, pp. 4610–4617.
- [182] Q. Wu, P. Dai, J. Chen, C.W. Lin, Y. Wu, F. Huang, R. Ji, Discover cross-modality nuances for visible-infrared person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2021, pp. 4330–4339.
- [183] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, J. Lai, Unsupervised person re-identification by soft multilabel learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 2148–2157.
- [184] Y. Lin, L. Xie, Y. Wu, C. Yan, Q. Tian, Unsupervised person re-identification via softened similarity learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2020, pp. 3390–3399.
- [185] D. Wang, S. Zhang, Unsupervised person re-identification via multi-label classification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2020, pp. 10981–10990.
- [186] Q. Yang, H. Yu, A. Wu, W. Zheng, Patch-based discriminative feature learning for unsupervised person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 3633–3642.
- [187] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, N. Zheng, Semi-supervised person re-identification using multi-view clustering, *Pattern Recogn.* 88 (2019) 285–297.
- [188] L. Qi, L. Wang, J. Huo, Y. Shi, Y. Gao, Progressive cross-camera soft-label learning for semi-supervised person re-identification, *IEEE Trans. Circuits Syst. Video Technol.* 30 (2020) 2815–2829.
- [189] H. Tang, Y. Zhao, H. Lu, Unsupervised person re-identification with iterative self-supervised domain adaptation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 2019, pp. 1536–1543.
- [190] G. Wu, X. Zhu, S. Gong, Tracklet self-supervised learning for unsupervised person re-identification, *Proceedings of the AAAI Conference on Artificial Intelligence* 2020, pp. 12362–12369.
- [191] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, T.S. Huang, Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2019, pp. 6112–6121.
- [192] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, Y. Tian, Ad-cluster: augmented discriminative clustering for domain adaptive person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2020, pp. 9021–9030.
- [193] X. Zhang, J. Cao, C. Shen, M. You, Self-training with progressive augmentation for unsupervised cross-domain person re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2019, pp. 8222–8231.
- [194] M. Ye, A.J. Ma, L. Zheng, J. Li, P.C. Yuen, Dynamic label graph matching for unsupervised video re-identification, *Proceedings of the IEEE International Conference on Computer Vision* 2017, pp. 5142–5150.
- [195] J. Chen, Y. Wang, J. Qin, L. Liu, L. Shao, Fast person re-identification via cross-camera semantic binary transformation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017, pp. 3873–3882.
- [196] G.A. Wang, S. Gong, J. Cheng, Z. Hou, Faster person re-identification, *Proceedings of the European Conference on Computer Vision* 2020, pp. 275–292.
- [197] C. Zhao, Y. Tu, Z. Lai, F. Shen, H.T. Shen, D. Miao, Salience-guided iterative asymmetric mutual hashing for fast person re-identification, *IEEE Trans. Image Process.* 30 (2021) 7776–7789.
- [198] G. Wu, S. Gong, Decentralised Learning From Independent Multi-domain Labels for Person Re-identification, <https://arxiv.org/pdf/2006.04150.pdf> 2020 (arXiv preprint arXiv:2006.04150).
- [199] S. Sun, G. Wu, S. Gong, Decentralised Person re-identification With Selective Knowledge Aggregation, <https://arxiv.org/pdf/2110.11384.pdf> 2021 (arXiv preprint arXiv:2110.11384).
- [200] W. Zhuang, Y. Wen, X. Zhang, X. Gan, D. Yin, D. Zhou, S. Yi, Performance optimization of federated person re-identification via benchmark analysis, *Proceedings of the ACM International Conference on Multimedia* 2020, pp. 955–963.
- [201] L. Bedogni, S.K. Rumi, F.D. Salim, Modelling memory for individual re-identification in decentralised mobile contact tracing applications, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2021, pp. 1–21.
- [202] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017, pp. 3415–3424.
- [203] B. Munjal, S. Amin, F. Tombari, F. Galasso, Query-guided end-to-end person search, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2019, pp. 811–820.
- [204] S. Tang, M. Andriluka, B. Andres, B. Schiele, Multiple people tracking by lifted multicut and person re-identification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017, pp. 3539–3548.
- [205] Y. Hou, L. Zheng, Z. Wang, S. Wang, Locality Aware Appearance Metric for Multi-target Multi-camera Tracking, <https://arxiv.org/pdf/1911.12037.pdf> 2019 (arXiv preprint arXiv:1911.12037).