EURASIP Journal on Advances in Signal Processing

SURVEY Open Access

Deep video-based person re-identification (Deep Vid-ReID): comprehensive survey

Rana S. M. Saad[1*], Mona M. Moussa[1], Nemat S. Abdel-Kader[2], Hesham Farouk[1] and Samia Mashaly[1]

*Correspondence:
rana@eri.sci.eg

[1] Computers and Systems Department, Electronics Research Institute, Cairo, Egypt
[2] Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Cairo University, Giza, Egypt

Abstract

Person re-identification (ReID) aims to find the person of interest across multiple non-overlapping cameras. It is considered an essential step for person tracking applications which is vital for surveillance. Person ReID could be investigated either using image-based or video-based. Video-based person ReID is considered more discriminating and realistic than image-based ReID due to the massive information extracted for each person. Different deep-learning techniques have been used for video-based ReID. In this survey, recently published articles are reviewed according to video-based ReID system pipeline: deep features learning, deep metric learning, and deep learning approaches. The deep feature learning approaches are categorized into spatial and temporal approaches, while deep metric learning is divided into metric and metric learning approaches. The deep learning approaches are differentiated into: supervised, unsupervised, weakly-supervised, and one-shot learning. A detailed analysis is held for the architectures of the state-of-the-art deep learning approaches. And their performance on four benchmark datasets is compared.

Keywords: Video-based person re-identification, Person re-identification, Deep learning, Literature survey, Video surveillance, Review, Survey

1 Introduction

Person re-identification (ReID) aims to identify persons through non-overlapping cameras. Identifying a person is to determine where and when the person appears in the recorded video set. Person re-identification grasps attention in the last decade due to its importance in surveillance applications, tracking, security monitoring, criminal investigation, finding lost people in mall centers, and forensic investigation [1].

The surveillance cameras widespread everywhere may indicate that identifying persons is a trivial task. However, person ReID systems face several challenges, some related to the person's appearance (as pose), low-quality camera resolution, and illumination changes. Other challenges are related to the surrounding effects such as occlusion, fusion, and background cluttering, while others are related to the bounding box misalignment [2].

1.1 Video-based person re-identification

Based on the input query, person ReID can be categorized into: image-based person ReID, video-based person ReID, and image-to-video ReID [3]. Image-based ReID is widely discussed in previous publications, while video-based is less served than image-based. In addition, this is the first review article dedicated to the deep video-based ReID. In a video-based ReID system, the input probe/query is a tracklet, multiple frames for the same person, while the output delivers the query into the gallery set.

Video-based person re-identification system consists of two major components: feature learning and metric learning. Feature learning aims to extract the intra- and inter-frame-level representative features. Intra-frame-level means learning features through frame content (local frame-level features) as its salient regions, quality, and resolution. However, the inter-frame-level features learn the associated features over frames, i.e. global video features [4]. Metric learning aims to measure the distance between the probe and the gallery set in order to find the best matched person.

Figure 1 clarifies the deep video-based ReID (Deep Vid-ReID) block diagram. Deep Vid-ReID pipeline takes the input tracklet (sequence of bounding boxes for the same person) and passes them to a deep learning model that learns both the spatial and temporal features for the input sequence. This model provides a representative information about salient information in the video. After that, a deep metric learning is introduced.

1.2 Review of state-of-the-art surveys

Several surveys have reviewed the person identification techniques such as [1–10]. Each of them addressed the re-identification task from a different point of view, combining the image-based and the video-based approaches in one survey. Up to our knowledge, there is no survey which has a dedicated review about video-based ReID only. For instance, Almasawa et al. [1] reviewed the person identification based on the input type as image-based, video-based, and image to video-based publications. A deep interest in the image-based approaches is discussed in their survey. They categorized the approaches into three main types: RGB-image, RGB-D image, and RGB-IR image. Compared to the discussion on image-based ReID approaches, video-based approaches are little served.

In addition, Wang et al. [5] discussed person ReID based on: (1) local information extraction from cameras as attention mechanism, (2) metric learning techniques; as

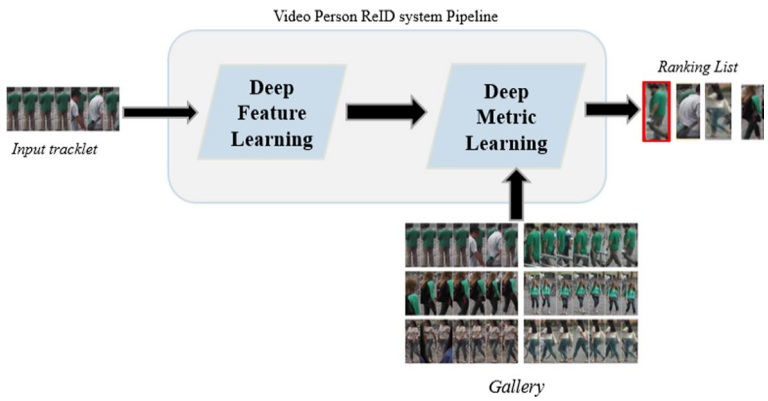

Fig. 1 Deep learning video-based person ReID system pipeline

distance learning, (3) data labeling problems, and (4) data types (video, depth, and IR ReID). The reviewed video-based approaches spotted mainly on the occlusion problems and redundant information within frames.

On the other hand, Leng et al. [6] focused on the open world scenario, and their review is based on the data procedure and efficiency. Besides Yaghoubi et al. [7] unified the categorization methods for the person ReID. A summary of different taxonomies was discussed based on the learning approaches, the identification settings, and the data types. The survey presented a general overview of person ReID.

Yadav et al. [8] reviewed the image-based ReID approaches based on architecture, challenges (pose, lighting, scale variation, and view variation), modality-based (either RGB or cross-modality as IR), and metric-based learning. On the other hand, the video-based ReID is discussed based on challenges and advantages. Only fourteen video-based publications are reviewed.

Ye et al. [2] presented the deep learning-based approaches from different perspectives, as open and closed world. Each perspective is discussed through certain points: deep feature learning, deep metric learning, and ranking methods. A new evaluation metric is introduced that reflects the cost of the correct matches. Video-based ReID is discussed as part of the closed world settings, where 17 video-based publications are only covered.

On the other side, Wu et al. [9] reviewed the deep learning-based approaches for re-identification but with different view. The survey organized the techniques into six types related to identification, verification, metric learning, part-based, video-based, and data augmentation. They addressed ten video-based publications.

In this literature, we aim to conduct a literature review to propose recent re-identification systems. The contributions of this survey are:

1. Presenting a literature review about the recent state-of-the-art methods for only video-based person re-identification for just deep learning models.
2. Presenting a comprehensive review of the datasets (either commonly used, benchmark datasets, or dedicated purpose), feature learning, and metric learning approaches.
3. Providing a comprehensive evaluation for each approach using a unified Rank 1 (R1) accuracy metric.
4. Presenting the challenges that face implementing person re-identification systems and recommendations to cope with these challenges.
5. Providing a detailed analysis of the state-of-the-art approaches over four benchmark datasets.
6. Introducing some trending applications and future directions that benefit from person re-identification.

Up to our knowledge, this is the first survey that reviews only the video-based person re-identification approaches based on the utilized deep learning architecture for the last 7 years. This survey differs from other surveys as follows:

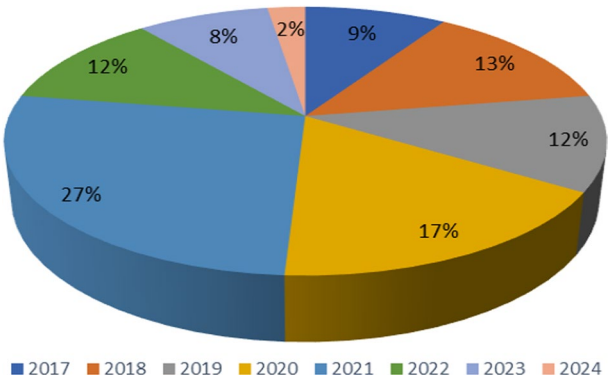

Fig. 2 Distribution of the number of publications from 2017 to 2024

Table 1 Number of reviewed papers in this survey

Year	2017	2018	2019	2020	2021	2022	2023	2024
No. of papers	15	22	19	28	44	19	14	4

1. Previous surveys discussed person ReID for both image and video approaches and did not provide a sufficient comprehensive review of video-based approaches. Little video-based publications were reviewed for previous surveys.
2. This survey covers the recent approaches for different scenarios as closed-world and open-world. It introduces different case studies such as cloth change and its related problems. On the contrary, previous surveys customized their survey for certain scenarios.
3. The survey is ordered to be a document for those interested in implementing video-based person ReID systems. It proposes the datasets at first, followed by deep features and metric learning sequentially as the implemented systems flow work.
4. The survey does not just review the deep learning-based approaches only, as shown in Fig. 3, but also gives a comprehensive analysis and comparative evaluation for each approach over four benchmark datasets in separate section.

1.3 Survey scope

This survey evaluates the recently published articles according to Rank-1 (R1) accuracy evaluation metric over four benchmark datasets: MARS, DukeMTMC-VideoReID, PRID2011, and iLIDS-VID. Figure 2 and Table 1 show the number of the published work till 2024 that will be discussed through the presented review.

The paper is organized as follows: The video-based person ReID datasets are illustrated in Sect. 2. The deep feature learning is stated in Sect. 3. Section 4 presents the deep metric learning. The deep learning-based approaches are discussed in Sect. 5, while discussion and future work are declared in Sect. 6. Figure 3 shows the survey structure.

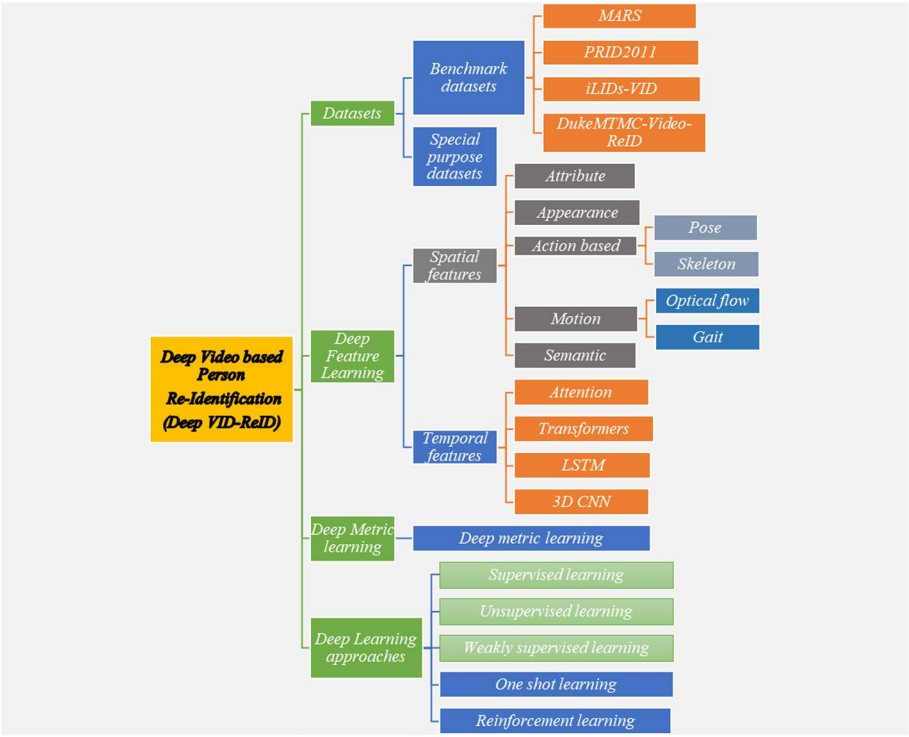

Fig. 3 The survey structure

2 Video-based person ReID datasets

In this literature, video-based ReID experiments are evaluated over some common benchmark datasets as MARS, PRID2011, iLIDS-VID, and DukeMTMC-VideoReID. Those datasets are covered through this survey, although some special-purpose datasets are also used for video-based ReIDs as: FGPR [11] and PoseTrackReID [12].

2.1 Benchmark datasets

MARS (Motion Analysis and Re-identification Set) [13] is considered the largest video-based person ReID dataset. It is an extension of the Market1501 dataset which is related to image-based ReID problems. The MARS dataset is constructed from 1261 pedestrians. Six synchronized cameras are used to construct the dataset. Each pedestrian is at least captured by two cameras. A Generalized Maximum Multi-Clique problem (GMMCP) tracker is used to automatically generate 20,478 video tracklets [14]. In addition, the dataset contains 3248 distractor sequences. MARS is evaluated by mean average precision mAP + and Cumulative Matching Characteristic (CMC) evaluation metrics. The dataset is publicly available [15]. A sample example is shown in Fig. 4.

PRID2011 The dataset images are gathered from two non-overlapping surveillance cameras [16]. 749 pedestrians are captured by one camera, and 385 pedestrians are captured by the other camera. Among those pedestrians, 200 persons appeared in both cameras. All images are cropped into 128×48 pixels. In clean and simple scene, the dataset images are captured with relatively illumination changes. PRID2011 is

Fig. 4 Sample example of *MARS* dataset. Credit goes to [13]

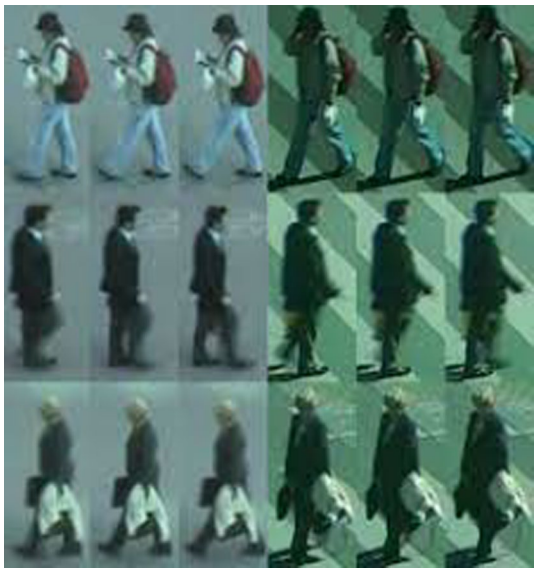

Fig. 5 Sample example of *PRID2011* dataset. Credit goes to [16]

evaluated using CMC evaluation metric. And it is available to download [17]. Sample example is shown in Fig. 5.

iLIDS-VID [18] is initially created from iLIDS Multiple-Camera Tracking Scenario (MCTS) dataset [19]. The iLIDS-VID consists of 600 videos from 300 pedestrians. Each pedestrian video is captured by two non-overlapping cameras fixed in an airport arrival hall. It is challenging due to lighting changes, and background occlusions. Video lengths range from 23 to 192 frame. iLIDS-VID is evaluated using CMC evaluation metric. Sample example is shown in Fig. 6.

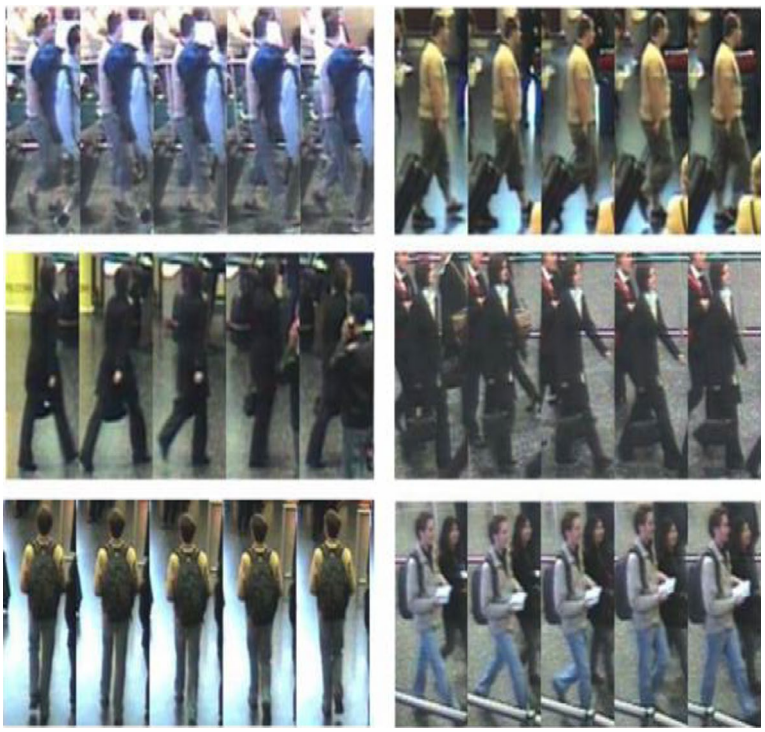

Fig. 6 Sample example of *iLIDS-VID* dataset. Credit goes to [18]

Fig. 7 Sample example of *DukeMTMC-VideoReID* dataset. Credit goes to [21]

DukeMTMC-VideoReID [20] Duke Multi-Target, Multi-Camera dataset is collected from Duke University campus to be used in person re-identification and multi-camera tracking research problems [21]. Eight synchronized cameras are used to capture the

students through the lectures break. Over 2700 pedestrian, two million images are captured. The dataset is evaluated using CMC metric. Sample example is shown in Fig. 7.

2.2 Special purpose datasets

Cloth changing is an important issue in person ReID specially in working areas with common uniform or for the different educational stages schools. It is important to retrieve the person of interest regardless the cloth they wear. For this purpose, some video-based datasets are collected as [11, 12, 22] as well as some image-based datasets as [23].

FGPR Fine-grained Person ReID dataset is a specific purpose dataset [11]. It is designed for person re-identification with relatively similar clothes such as a uniform. The appearance features are not inevitable to discriminate persons. Thus, pose-based features such as motion-attentive and joint-specific local dynamic pose features are proposed and robustly evaluated over FGPR. FGPR was collected from three groups (blue, green, and white) wearing the same clothes color. 358 pedestrians are captured with five cameras. 716 tracklets are formed and evaluated by CMC and mAP evaluation metrics. A sample example is shown in Fig. 8.

Deep change Long-term Person Re-identification: [22]. This dataset is categorized as a large-scale dataset. Seventeen cameras are used to capture 1082 identities gathered over 12 months. This long period of recording provides a realistic personal appearance with different weather conditions and different walking styles. mAP and CMC are used as evaluation metrics. The dataset and code are publicly available [24]. A sample example is shown in Fig. 9.

PoseTrackReID [25] It is a dataset constructed to solve some multi-person pose-tracking problems such as person appearance in multiple frames with different occlusion and obstacle views [12]. It is sampled from the original Pose ReID dataset [26] such that various sequences are sampled from the same video. Thus, occlusion and blurring are highly found.

Fig. 8 Sample of *FGPR* dataset. Credit goes to [11]

Fig. 9 Sample of *Deepchange* dataset. Credit goes to [22]

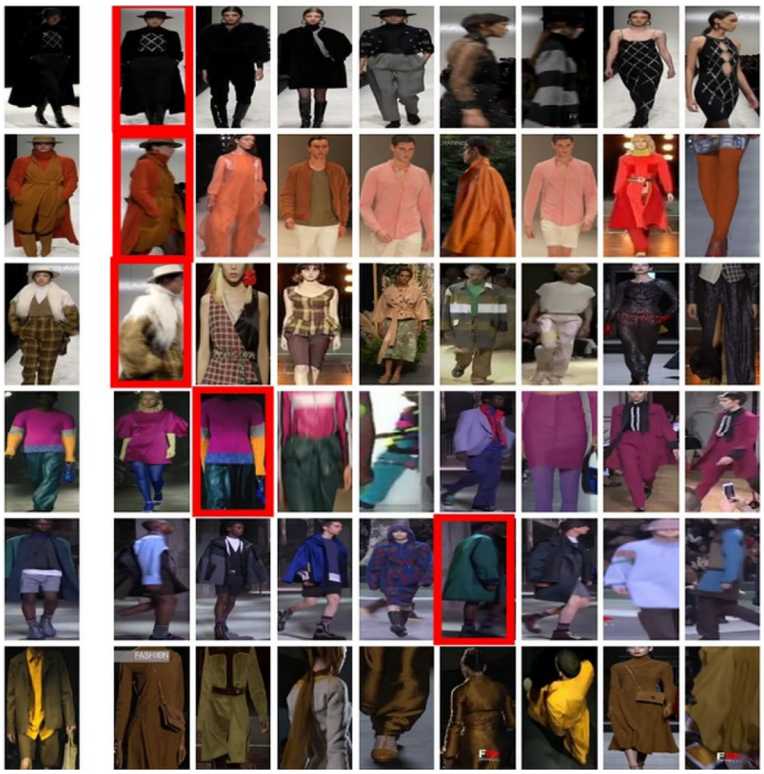

Fig. 10 Sample of SYSU-30k dataset. Credit goes to [27]

SYSU-30k Wang et al. [27] introduced the first large publicly available dataset, SYSU-30k. The original videos are downloaded from the internet TV programs. It is about 1000 raw videos. Annotators have collected 30,000 identities with weak supervision style with 84,930 bags. A sample example is shown in Fig. 10.

BUAA-Duke-Gait dataset As most Gait-based datasets, it is constructed under some rigid constraints, such that the persons have to walk in a straight line with no one appearing in the camera view. These conditions are not realistic for the common surveillance system. Thus, Shaoxiong Zhang et al. [28] constructed the BUAA-Duke-Gait dataset. It is deducted from the original DukeMTMC-VideoReID dataset by extracting the binary silhouettes of the colored pedestrian appears in the original dataset. Thus, BUAA-Duke-Gait dataset is constructed with the aid of eight cameras with total 4,612 video and 3,623,488 frames. Figure 11 shows a sample of the dataset.

Fig. 11 The BUAA-Duke-Gait dataset generation from the original DukeMTMC-VideoReID dataset. Credit goes to [28]

Fig. 12 The MEVID dataset. Credit goes to [29]

Davila et al. [29] constructed the Multiview Extended videos with Identity (MEVID) dataset. It is an extended version of the MEVA dataset [30]. MEVID is constructed to give wide variation for indoor/outdoor scenarios. Over 73 days and different camera views, 590 frames were captured using 33 cameras f 158 persons. Different wearing outfits form 8092 tracklets. MEVID is annotated using a semi-annotated tool designed from different models to handle object detection and tracking with various poses. Figure 12 shows a sample of the dataset.

3 Feature-based learning

Although videos provide richer information about pedestrians than images, there are much more redundant information in the frames sequence. Nevertheless, it is useful to find a method that captures the leveraging discriminative features and discard others.

Video-based person ReID features are classified according to the feature representation level into frame level and video level, also called local and global, respectively. For frame level, spatial features are acquired, and temporal features are determined and aggregated to get the global level representation. In this survey, the spatial feature is categorized based on its type into attribute, pose, motion, and appearance. On the other hand, the temporal features are categorized into attention, transformers, LSTM, and 3D CNN-based approaches.

The performance of the proposed techniques is evaluated using the R1 score which is related to the top-ranked query results from the search in the gallery set [31].

3.1 Spatial Features

3.1.1 Attribute-based approaches

Zhao et al. [32] used an attribute-based method for weighting each frame according to its dominating attribute. Each person can be described by some attributes that define him. They formulated the attributes into 6 groups: gender, head, shoes, upper body, lower body, and attachments. For each attribute, ResNet-50 is used to extract the features from it. After that frames across the sequence are reweighted, such that higher weights are assigned to frames with the most dominating attributes. This helps in giving an effective representation of the detected person. On MARS dataset, 87% R1 accuracy is achieved.

Xu et al. [33] acquired the spatial–temporal information from the video sequences. For each frame region, Siamese CNN (SCNN) network is utilized to extract the spatial information. SCNN also works on learning the frame-level representation over frames in the sequence, while the temporal relation between frames over the sequence level is obtained by RNN and the distance metric between frames. An attention model is then used to jointly represent all changes through a sequence of frames. The results show 62%, 77%, and 44% R1 accuracy for iLIDS-VID, PRID2011, and MARS datasets, respectively.

Based on cameras low-resolution capabilities and limited availability of labeling datasets, Zhang [34] built their model to solve these difficulties. They used CNN to detect the attribute-based features such as head, torso, and legs. Then, several mining rules are used to refine those features. The results are combined with XQDA metric learning to convert the attribute-based classification into a re-identification task. An appearance model [35] is also attached to finalize the re-identification system. PEdesTrian Attribute (PETA) dataset is used for attribute training, while iLIDS-VID and PRID2011 are used for validation and testing. 60.3% and 73.2% R1 ranking accuracy are achieved.

Yin et al. [36] grabbed the attention toward the problem of person ReID with similar clothes, i.e. those have the same uniform. Those people usually have the same appearance and the same cloth color, and thus, the dependency on appearance as discriminative features is disappointing. For this purpose, Yin et al. defined this case as fine-grained ReID and proposed two dynamic pose features to solve it: motion-attentive and joint-specific local dynamic pose features. The motion-attentive features aim to generate masks on important areas in each frame using RNN, while the joint-specific pose features aim to estimate the pose for each body region per frame using CNN. The proposed model achieved 87.1% R1 accuracy for their self-built dataset and 82.9% over MARS dataset.

Song et al. [37] used multi-task CNN to learn nine deterministic attributes for feature extraction. Some of the attributes are stationary attributes such as upper body, lower body, hair, and gender, while the bag is used as a dynamic attribute. In training, the nine attributes for each person are learned through CNN and then get more robust by using the local maximal co-occurrence (LOMO) descriptor [38]. long short-term memory (LSTM) is used then for frame aggregation. On MARs, 85.8% R1 accuracy is achieved while 95.6% and 83.9% for PRID2011 and iLIDS-VID, respectively.

Chen et al. [39] have used the Attribute-aware Identity hard Triplet Loss (AITL) to solve the Variance among Different Positives (VDP) problems. Although the triplet loss discards the anchor from the negative samples, there are some similarities between the positive samples that have to be reduced for best-ranking results. With the aid of the spatial–temporal attribute-driven attention model, the VDP problem is decreased. ID relevant attributes as bottom and top color and ID irrelevant attributes as occlusion, pose, and motion were used. 88.2% R1 accuracy results on MARs is achieved.

Song et al. [40] introduced a Two-Stage Attribute Constraints (TSAC) network. The two-stage network helps in extracting the image level features and the sequence-level features using CNN and LSTM, respectively. The image-based stage is introduced for four static attributes: upper body, lower body, hair, and gender. The fifth dynamic attribute, bag, is introduced through the sequence-level features as a discriminating feature. TSAC attained 84.1%, 64.3%, and 52.6% R1 classification accuracy over PRID2011, iLIDS-VID, and MARs datasets, respectively.

Chai et al. [41] proposed Attribute Salience-Assisted Network (ASA-Network) to solve background and pose variance issues. The ASA-Net is composed of five branches to learn the ID-relevant attributes such as hair, gender, and upper body cloth and the ID-irrelevant attributes such as pose and motion. Five different metric learnings are used in ASA-Net such as triplet loss, cross-entropy, center loss, binary cross-entropy, and pose and motion invariant loss. ASA-Net accomplished 90.2% and 97.6% R1 accuracy results over MARS and DukeMTMC-Video ReID datasets, respectively.

Multiple feature fusion Network (MPFF-Net) is proposed by Song et al. [42]. Both hand-crafted features and deeply learned features are introduced. The LOMO descriptor is used as hand-crafted features which is for frame-level feature learning. Then, bilinear LSTM (Bi-LSTM) deeply aggregated those features in forward and backward directions. They tested MPFF over iLIDS-VID, PRID2011, and MARs datasets. R1 accuracies results are: 88.1%, 73.1%, and 84.5%, respectively.

Using six attribute groups, Zhao et al. [32] described each person depending on these attribute groups and reweighted each attribute according to its appearance in the frame. The six attribute groups are related to gender, head and shoulder, upper body, lower body, shoes, and attachments. Zhao, Yiru, et al. get the benefits of transfer learning from the attribute recognition dataset (RAP) [43] into the ReID problem instead of supervised learning. The built model resulted in achieving 82.6%, 91.7%, and 81.5% R1 accuracy over MARS, PRID2011, and ILIDs-VID datasets, respectively.

Chen et al. [44] have annotated the MARs dataset into 16 attributes. The attributes are divided into identity-relevant (such as gender, bottom, and carrying a handbag) and behavior-relevant (pose and motion). A multichannel network is built using CNN and temporal attention. The network is trained on the annotated dataset. The model attained

Table 2 The attribute-based approaches (R1 accuracy in %)

Authors	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Zhao et al. [32]	Attribute-based	87	–	–	–
Xu et al. [33]		44	–	77	62
Zhang et al. [34]		–	–	60.3	73.2
Yin et al. [36]		82.9	–	–	–
Song et al. [37]		85.8	–	**95.6**	83.9
Chen et al. [39]		88.2		–	–
Song et al. [40]		52.6	–	84.1	64.3
Chai et al. [41]		**90.2**	**97.6**	–	–
Song et al. [42]		84.5	–	73.1	**88.1**
Zhao et al. [32]		82.6	–	91.7	81.5
Chen et al. [44]		87.01	89.31	–	–

The bold values indicate the best R1 accuracy results

87.01% R1 accuracy on MARs, while 89.31% R1 accuracy on DukeMTMC-VideoReID dataset. The attribute-based approaches are summarized in Table 2.

For the attribute-based approaches, it is noticed that composing the ID relevant attributes with ID irrelevant ones such as motion and pose achieves better results as in [41, 44]. The ASA-Net [41] outperforms other high results by 2% over MARS and more than 8% over DukeMTMC-VideoReID. Their approach was not tested over PRID2011 and iLIDS-VID datasets. On the other hand, only stationary ID attributes system results could be enhanced by using temporal learning architecture as LSTM [37]. In [37], competing results on PRID2011 are acquired by about 4% R1 accuracy. The combination of hand-crafted features at spatial and temporal levels did not introduce better results as in [42]. It is recommended in the future work to try another temporal feature discriminator such as the 3D CNN over the attribute-based features and evaluate the results for the ID-attribute features. Furthermore, it is also recommended to combine both ID and non-ID attribute features and introduce them to robust temporal LSTM, 3D CNN, and attention models for achieving better results.

3.1.2 Appearance-based approaches

Appearance is one of the commonly used features for person ReID. It has been addressed for image- and video-based systems. Here, a video-based person ReID system would be reviewed.

Zhang et al. in [45] used multiple CNNs to extract the salient appearance features from certain representative frames instead of the whole sequence. Depending on the person's walking profile, the minima and maxima of the flow energy profile (FEP) signal [46] are exploited in selecting the salient frames. Then, multiple CNNs followed by pooling layers are used to extract features from the selected frames. Those features are further passed into metric learning for re-identification. Over MARS, iLIDS-VID, PRID2011, and SDU-VID dataset, Zhang et al. got 55.5%, 60.2%, 83.3%, and 89.3% R1 accuracy, respectively.

For sequential frames, different person poses may lead to different bounding box sizes which causes misalignment in appearance feature learning for the convolution process in CNN and hence the appearance features will be destructed. For this purpose, Gu

et al. [47] proposed the appearance-preserving 3D convolution (AP3D) framework. The AP3D has an appearance-preserving module with a 3d convolution kernel that works on learning the adjacent and central features map between two frames. Over MARS, DukeMTMC-VideoReID, and iLIDS-VID, 90.7%, 97.2%, and 88.7% R1 accuracy is achieved, respectively.

3D CNN is used to learn the temporal and spatial cues in video-based ReID. However, due to its large number of parameters, Li et al. [48] added multi-scale 3D layers (M3D) that have different temporal ranges to reduce the number of training and optimization parameters. A two-stream network is proposed for spatial and temporal features learning, and the M3D is only presented in the temporal stream. 84.3%, 94.4%, and 74% R1 accuracy is achieved for MARS, PRID2011, and iLIDS-VID datasets, respectively.

Although the M3D layers in [48] improved the results, Li et al. [25] have modified the proposed M3D architecture to define local and global M3D. The local M3D is introduced as in [48] in order to acquire the spatial–temporal frame features, while the global M3D is proposed to avoid the misalignment between the adjacent frames. Experimental results show that the R1 accuracy for MARS, PRID2011, and iLIDS-VID is improved to reach 86.3%, 96.6%, and 86.6% for each, respectively.

The global features do not capture the relation between adjacent frames only, and they are also useful for reducing the occlusion effect through the whole sequence as in [49]. Li et al. have proposed the global features in two ways: short term and long term. The short-term global features are implemented by dilated convolution to get the appearance features of the adjacent frames, while the long-term are implemented by the self-attention model for inconsecutive frames. Results show that the R1 accuracy is 87.02% for MARS, 94.6% for PRID2011, and 96.29% for DukeMTMC-VideoReID.

The frame misalignment is counted as a problematic issue for video-based ReID. The non-local block is used as a part of the framework in [50]. Liao et al. [50] used 3D CNN to represent the spatial–temporal relationship through frames; however, the non-local block is used for misalignment problem and long-term video sequence relations. The experimental results show 84.3%, 91.2%, and 81.3% R1 accuracy on MARS, PRID2011, and iLIDS-VID datasets, respectively.

To conclude, appearance-based features play a key role in representing video spatial features. It is noticed that the deep appearance feature achieves high accuracy. However, appearance may be affected by sequence misalignment. Thus, AP3D in [47] has an efficient appearance feature map that saves the appearance and reduces the misalignment effect more than the traditional CNN proposed in [45] for appearance representation. In addition, supporting the appearance-based features with a deep-based temporal feature backbone enhances the results. AP3D in [47] outperformed [45] work by more than 35% and 22% over MARS and iILDS-VID, respectively, by applying 3D CNN for the deep temporal guidance. It is recommended in the future researches to propose an experiment that constructs 2D, and 3D graph maps then learn the appearance features over them. The appearance-based approaches are summarized in Table 3.

3.1.3 Action-based approaches

Action-based approaches depend on the person's motion action. It is addressed in two manners: using the pose information for each tracklet or using the skeleton structure.

Table 3 The appearance-based approaches (R1 accuracy in %)

Author	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Zhang et al. [45]	Appearance-based	55.5	–	83.3	60.2
Gu et al. [47]		**90.7**	**97.2**	–	**88.7**
Li et al. [48]		84.3	94.4		74
Li et al. [25]		86.3	–	**96.6**	86.6
Li et al. [49]		87.02	96.29	94.6	–
Liao et al. [50]		84.3		91.2	81.3

The bold values indicate the best R1 accuracy results

Pose is one of the discriminative features for video-based person ReID. It is widely used in both image-based ReID [51–54] and video-based ReID [55–58].

The pose could be estimated coarsely or finely [8, 59]. In coarse pose estimation, the pose is estimated according to camera parameters such as calibration settings, ground plane position, 3D position, and person velocity as in [45, 46]. Fine estimation corresponds to person joint key points that were recently estimated by deep-based models such as deep cut [60], pose-driven deep convolutional (PDC) [58], and pose-invariant embedding (PIE) [61]. Both coarse and fine pose estimation is integrated in many frameworks to solve the body parts misalignment problem [61].

For that purpose, Wei et al. [55] built a posture-based video ReID in two sequential stages: pose alignment module and multi-scale structure learning module. The pose alignment module extracts a representative local frame-level feature by selecting the best posture frames. These frames are selected by comparing the reference pose set and the estimated pose for each frame. The pose feature according to each body region is hence extracted. Moreover, the multi-scale structure learning module exploits the relationship among those regions features by using a graph convolution network (GCN). MARS, iLIDS-VID, and PRID2011 datasets are used for model evaluation. 90.2%, 85.5%, and 94.7% R1 accuracies are achieved for each dataset, respectively.

Gao et al. [56] also proposed pose estimation as an efficient video sequence alignment tool. The distance between ankles is used to define the pose, while a set of poses is defined as reference poses. Hence, a temporal representation is achieved. On the other hand, the body regions are defined as spatial features. Cross-view quadratic discriminant analysis (XQDA) is used as a distance metric. 92.9% and 74.6% R1 accuracies regarding PRID2011 and iLIDS-VID datasets are accomplished, respectively.

As a solution for the frames misalignment problem, the coarse and fine pose was introduced by Sarfraz et al. [52]. According to the camera view, the coarse pose is defined as front, back, and side. For each view, a three-part CNN is used to extract the pose information to propose a robust representation. Also, for each view, a fine pose belonging to fourteen body joints is used by applying a deeper cut module [60]. In a plus, a new re-ranking technique based on expanded cross-neighborhood is proposed for increasing the retrieval performance. The system is tested on image and video-based datasets. 64.6% R1 accuracy is achieved on MARS.

The skeleton information is a good motion guidance. Elaoud et al. [62] used the person skeleton as input to the tracklets. The shape of the skeleton is projected on the

Grassmann manifold [63], and each person's camera view is weighted dynamically according to its joint distance on Grassmann. Those weights are classified by random forest. Results on iLIDS-VID and PRID2011 show 58.06% and 87.02% R1 accuracy, respectively.

Similarly, posture structure and action constraints—Hypergraph Pedestrian Video Re-identification (PA-HPAReid), are proposed by Hu et al. [64] to avoid pose change in re-identification. The color and posture structures are estimated by CNN with the aid of the pedestrian skeleton estimation method [65]. Then, GCN is used as a graph modeling for the estimated skeleton joints. Relative entropy loss is used as metric learning for the salient regions. On MARS, iLIDS-VID, and PRID2011, PA-HPAReid attained 89.9%, 87.9%, and 95.9% R1 accuracy, respectively.

Spatial–temporal Correlation and Topology Learning CTL is proposed by Liu et al. [66]. In CTL, for the input video sequence, the person's key points are estimated and then categorized into three scales as multi-scale extraction. For each scale, the 3D graph is constructed, and the local features are extracted from it. These scales are then fused to extract the global feature. Experiments on MARS and iLIDS-VID show 91.4% and 89.7% R1 accuracies, respectively.

Wei et al. [67] fused the action-based features with the appearance-based features through 3D CNN. The action-based information is gathered in a pyramidal manner using three channels: R, G, and B. Each channel uses five consecutive frames to attain the motion information in two steps by recording the change in motion between frames. Further triplet loss is introduced for the output of the fused appearance and action network result. Over MARS, iLIDS-VID, and PRID2011, 86.7%, 86.5%, and 94.6% R1 accuracies are accomplished, respectively.

Graph-based solutions have recently been introduced for image- and video-based ReID for different purposes such as noisy label reduction as in [68] and in [69] to explore the correlation between body parts as in Lu et al. work [69]. The correlation is determined using a dynamic hypergraph over the temporal skeletal information. A mix of joint-centered, bone-centered hypergraphs, and multi-granularity spatial–temporal information are presented as a framework for better feature representation. Over iLIDS-VID, PRID2011, and MARS, 92%, 96.9%, and 92.5% R1 accuracies are accomplished.

For action-based features, it is noticed that the graph representation is suitable for improving the posture points. The 3D graph structure in Liu et al. [66] has a competing result over others as in [57, 64]. There is about 1.5% and 1.8% improvement over MARS and iLIDS-VID, respectively. PRID2011 is not tested by [66], while the competing [57, 66] has a test over PRID2011 achieving 95% R1 results. Although 3D CNN achieved less in [67], it is recommended to use it with graph construction based on [60, 66] work. Results for the action-based approaches are summarized in Table 4.

3.1.4 Motion-based approaches

3.1.4.1 Gait-based approaches Person gait is categorized as one of the human soft biometrics [65–72]. Gait recognition for person re-identification depends on the pedestrian's walking style as the gait has a unique pattern for the pedestrian movement. Fortunately, gait has some important characteristics: It is unique and can be measured for any distance [70]. Unfortunately, it is affected by aging and illness. Nambiar et al. [65–72]

Table 4 The action-based approaches (R1 accuracy in %)

Author	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Sarfraz et al. [52]	Action-based	64.6	–	–	–
Wei et al. [55]		90.2	–	94.7	85.5
Gao et al. [56]		–	–	92.9	74.6
Elaoud et al. [62]		–	–	87.02	58.06
Hu et al. [64]		89.9	–	**95.9**	87.9
Liu et al. [66]		**91.4**	–	–	**89.7**
Wei et al. [67]		86.7		94.6	86.5

summarized the gait-based approaches in re-identification and categorized them into five categories based on: (1) camera settings, (2) pose and gait direction (either pose-dependent or pose-independent), (3) feature extraction model (model-based and model-free), (4) classification approach (learning and non-learning approaches), and (5) application scenario (either spatial or temporal approaches).

Other work by Sepas-Moghaddam et al. in [71] summarized the deep gait recognition approaches. They noticed the transition in the last 6 years from non-deep-based approaches into deep-based. Sepas-Moghaddam categorized the deep-based approaches into some aspects as body representation, temporal representation, feature representation, and finally according to nine different deep architectures.

For skeleton-based datasets (as CASIA [72], OU-ISIR [73], and OU-MVLP [74]), Elharrouss, Omar, et al. [75] used two CNN models for gait-based ReID. The input frames are imported to background and segmentation models to separate the skeleton from frames, and then, the Gait Energy Image (GEI) is extracted by the first CNN. GEI helps in estimating the gait angle. The second CNN helps in verifying the gait angle. Rao et al. [76] introduced self-supervised learning to learn the gait from the unlabeled skeleton. The self-learned model could reconstruct the unlabeled skeleton, and then, a locality-aware attention mechanism is introduced to save the video inter- and intra-skeleton sequence and get the final gait encoding feature representation.

Zhang et al. [28] proposed an effective gait recognition method that works under the color silhouettes and has 94% R1 accuracy for the ReID task compared to the binary silhouettes which have 78.24% R1 accuracy. The proposed gait recognition method consists of four modules: random tracklet sampling, alignment network, backbone network, and patch pyramid mapping. The experiments are held on the BUAA-DukeGait dataset which is constructed specially for realistic gait-based scenarios.

A recent published review by Rahi et al. [77] reviewed the gait-based person ReID and covered old and recent approaches for image- and video-based. They do not focus on the video-based approaches.

Instead of working on RGB frames, Zhao et al. [78] used the silhouette mask for frames as a discriminative gait sequence that has rich motion and appearance cues. Then, an attention module is presented to the appearance and gait features. Results on MARS and DukeMTMC-VideoReID have 84.3% and 96% R1 accuracies, respectively.

The frames masking is also presented by Chang et al. [79]. They also exploit the appearance and gait-based features. The masking is computed as foreground to

background portion, which covers at least 15% of the foreground. They called the masked dataset "Mask-MARS." A gait set using CNN is proposed as a temporal discriminator. R1 results on Mask-MARS achieved 87.3%.

To conclude, the gait-based approaches are massively evaluated for different datasets especially for the cloth change cases as in [80–82]. The idea is to extract the frames appearance and GEI and then exploit the temporal features with an appropriate discriminator. To achieve better results in the future, it is recommended to use LSTM or 3D CNN or attention model as a temporal feature backbone [78]. In addition, it is interesting to try the 3D shape person modeling instead of silhouettes for person representation and then propose the resulting model over the four-benchmark datasets.

3.1.4.2 Optical flow-based approaches Optical flow defines the object's motion through consecutive frames within a sequence. It is determined by the relative distance between the object and the camera [83]. The optical flow is used earlier as a representative feature of video tracklets as in [84]. McLaughlin et al. in [84] used color and optical flow as input features to their network. Color feature represents the appearance metric, while optical flow represents the motion metric. Afterward, CNN, RNN, and temporal pooling are introduced as complete feature representation for the whole sequence. Siamese network is used then for training. Experiments over iLIDS-VID and PRID2011 datasets show 58% and 70% R1 accuracies, respectively.

In addition to its motion variation detection capability, optical flow map works as an appearance mask for persons in each frame [85]. For this purpose, Kiran et al. [85] inspired his work from Cho et al. [86] and merged the optical flow feature with the salient appearance feature for long-term temporal feature aggregation. The attention network exploits the salient appearance frame features for both the original frames and its corresponding optical flow map. Then, weighted feature aggregation weights the most salience frames for the final representation. Experiments over MARS, DukeMTMC-VideoReID, and iLIDS-VID show 86.6%, 96.7%, and 88.1% R1 accuracy, respectively.

In Chung et al. [87], two separate Siamese networks are exploited to learn the spatial–temporal co-occurrence features from both the raw RGB and the optical flow map frames separately. Each network representation later is proposed to the weighted features. Those features compare the Siamese cost for both branches and choose the best. Experiments over iLIDS-VID and PRID2011 show 60% and 78% R1 accuracy, respectively.

Chen et al. [88] proposed a reinforcement learning A3D attention network. The network learns the spatial temporal features from the 3 bins: the raw RGB frames, optical flow map frames, and the RNN temporal representation. The attention model is used to learn the salient features of each bin. The resulting representation selection is formulated in the Markov model and optimized by reinforcement learning. R1 accuracy results show 95.1%, 87.9%, and 86.3% for PRID2011, iLIDS-VID, and MARS, respectively.

The optical flow maps and the raw RGB frames are concatenated into a single channel and fed as an input to a competitive snippet-similarity aggregation network proposed by Chen et al. [89]. For both gallery and probe sequences, short length snippets are formed and aggregated from the original frames with the guidance of deep co-attention similarity model. The resulting snippets have less intra-frame variation, and thus, when the

probe snippet is requested, simple matching with similar gallery snippets is performed. The R1 accuracy results for iLIDS-VID, PRID2011, and MARS are: 85.4%, 93.0%, and 86.3%, respectively.

To conclude, it is noticed that combining the optical flow map with the appearance feature obtains better results as in [85]. Around 0.3% improvement on MARS and 0.2% on iLIDS-VID proves that the proposed systems [85, 88] are competing. For better results, it is recommended to do frame weighting as in [85] and apply it to the A3D architecture proposed in [88]. The optical flow approaches are summarized in Table 5.

3.1.5 Semantic-based approaches

Hou et al. [90] also work on global spatial–temporal context. Interaction–Aggregation-Update (IAU) network for image and video ReID is proposed. The upper body part is noticed to hold the more discriminative features for the person. Therefore, two networks are constructed: spatial–temporal IAU (STIAU), which aggregate the body part local features by CNN, and a channel IAU that forms the temporal features for the semantic body parts across all frames. The interaction between STIAU and CIAU is modeled. 90.2% R1 accuracy is achieved on MARS and 89.3% over DukeMTMC-VideoReID.

3.2 Temporal Features

3.2.1 Attention-based approaches

Attention in images means focusing on the important /salient parts of the image. Thus, attention-based methods could be used to pick the local and significant features in a given Bounding Box, (BB) [91]. To extract the strict BB information, deep learning-based approaches are used to solve the misalignment problem [92]. Recently, an attention mechanism has been proposed to attain the ReID task. [92–101] performed image-based person ReID and tested their approaches on Market1501, DukeMTMC-ReID, CUHK03-NP, and MSMT17 datasets. For example, JWASS in [102] eliminated the non-important background (as the common background between various persons), which resulted in reducing the model performance.

For this reason, Ning et al. in [102] proposed the attention mechanism for the background removal of non-salient regions and focus on the person's salient region features. The attention framework in their model consists of two branches: The first is the backbone ResNet with a background salience module, and the second branch is the attention-aware module which has an attention generator, attention map, and feature fusion

Table 5 The optical flow-based approaches (R1 accuracy in %)

Author	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
McLaughlin et al. [84]	Optical flow-based	–	–	70	58
Kiran et al. [85]		**86.6**	**96.7**	–	**88.1**
Chung et al. [87]		–	–	78	60
Chen et al. [88]		86.3	–	**95.1**	87.9
Chen et al. [89]		86.3	–	93.0	85.4

The bold values indicate the best R1 accuracy results

to form the final representative feature. Their work was tested on Market1501 dataset and achieved 96.8% R1 accuracy.

Attention models were also applied for video-based person ReID to discover salient regions [103], to select important frames and clips [104–106], and to learn discriminative representations [33, 106, 107].

Chen et al. [88] noticed that many spatial–temporal attention methods extract the salient regions in each frame distinctly. They worked on solving this independence of traditional salient attention models. Thus, a framework of multiple stages is constructed. The first stage extracts the appearance and optical flow stream features and aggregates them over frames to propose global representation. The second stage uses RNN to learn the location of salient attention regions in frames. Reinforcement learning is used to optimize the most representative frames. Over iLIDS-VID, PRID2011, and MARS, the framework is evaluated and accomplished 87.8%, 95.1%, and 86.1% R1 accuracy, respectively.

Shu et al. in [108] extracted the important discriminative cues from frames as shoes, hair, and bag using the attention mechanism. Additional cues (that extract the appearance and the color) are detected beside the global attention. Multiple spatial and temporal attention models were used. The spatial attention model split the feature map into channels and got the attention weights for each. R1 accuracy results on MARS are raised to 89.0% rather than 83.7% for COSAM [109].

Wu et al. [110] used the attention network to learn the salient regions in frames. The spatiotemporal representation is described by their model. The model consists of CNN that learns the spatial features for each frame, and then, the Siamese network with GRUs and attention model explore the salient regions and neglect other non-salient regions. The R1 accuracy results are 61.2%, 74.8%, and 69.7% over ILIDS-VID, PRID2011, and MARS, respectively.

The occlusion and pose variation problems are reduced by Fu et al. [106]. Clip representation for each person is constructed, and salient frames are extracted. The frames are proposed to CNN to extract the spatial features followed by the attention module. An attention score is assigned for each frame to form the attention score matrix. Frames with high attention scores mean that it has a salient region. The fusion between salient frames forms the final clip representation. The proposed method R1 accuracy results are 86.3% and 96.2% for MARS and DukeMTMC-VideoReID, respectively.

Zhang et al. [111] proposed an attention module for multiple granularities. Granularity is defined as regions with different sizes [111]. The spatial features are extracted by ResNet-50, and then, attention scores for the set of reference feature nodes (S-RFN) are generated. R1 accuracy results are 88.8%, 88.6%, and 95.9% for MARS, iLIDS-VID, and PRID2011, respectively.

Although several approaches were proposed for temporal attention, Chen et al. [112] aimed to get more accurate sequence-level features for the salient frames. Thus, a dual-constrained guided network (DCGN) is proposed. DCGN is composed of two stages: The first is for frame-level feature representation with optimal frame selection strategy to select the best representative frame from the sequence employing CNN and Frame-Constrained Module (FCM). The second stage is for temporal modeling by assigning the attention weights over the whole video to select the linked frames

using Sequence-Constrained Module (SCM). In a plus, an optimizing module is also introduced for classification and metric loss. R1 accuracy results are 89.6%, 95.35%, 78.51%, and 90.82% for MARS, DukeMTMC-VideoReID, iLIDS-VID, and PRID2011, respectively.

To remove the frame redundancy, while avoiding important information loss in representative frames, Jiang et al. [4] performed feature fusion over multiple temporal attention on different semantic levels. The input tracklet is introduced to different semantic levels. For each semantic level, an intra-frame temporal attention network is defined. These networks preserve all salient information. Inter-frame temporal attention is also performed to remove redundancy between frames. Finally, the various temporal attentions of the salient levels are fused to get the final representation of the input tracklet. Results on MARS, PRID2011, and iLIDS-VID show 87.1%, 95.8%, and 87.7%, respectively.

Yang et al. [113] used the attention model to get a relation between the spatial–temporal global and partial features. The global features get a relation map across frames while the partial features work on getting a relation between frames for the same spatial position. 89.1%, 97.2%, and 88.7% are the R1 accuracy results for MARS, DukeMTMC-VideoReID, and ILIDS-VID, respectively.

Bayoumi et al. [114] proposed a pyramid multi-part attention model (PMP) with multi-attention (MA) model. The PMP part aggregated features with three different details levels according to the person captured part. Each level summarized the person's details and introduced them to the MA part. The MA part extracted the salient features for each level. Results on MARS, DukeMTMC-VideoReID, PRID2011, and iLIDS-VID have 90.6%, 97.2%, 98.9%, and 92.8%, respectively.

In Wang et al. work [115], information channels for each frame are identified and concerning the whole sequence utilizing attention networks. This improves the spatiotemporal information representation and the correlation between frames. Even though the proposed frame weighting module not only weighted the individual frames but also corresponded to the sequence. Over MARS, DukeMTMC-VideoReID, and iLIDS-VID; the proposed work achieved 90.4%, 97.7%, and 90% R1 accuracy, respectively.

Tao et al. [116] presented an interference and pixel noise removal framework by using the attention mechanism. The framework mainly consisted of two basic modules: one for removing the interfering frames and the other for removing the noisy background that over-interfered the representative feature. An additional adaptive mask generated by the augmentation technique is introduced for further smoothing of the selected person. Over Mars, the framework outperformed 92.5% Rank 1 accuracy.

Bai et al. [117] introduced the broad idea as an extension to the salient feature for each frame. The subsequent frames have the same information that exists in the last one but with additional information. This paper suppressed the replica information obtained by the attention module and then kept the additional one. 91.0% and 86.7% R1 accuracies are achieved. The attention-based approaches are summarized in Table 6.

3.2.2 Transformer-based approaches

Transformers are initially presented for natural language processing applications as translation [118–120]. In recent years, transformers have been transplanted for different

Table 6 The attention-based approaches (R1 accuracy in %)

Author	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Chen et al. [88]	Attention-based	86.1	–	95.1	87.8
Shu et al. [108]		89.0	–	–	–
Wu et al. [110]		69.7	–	74.8	61.2
Fu et al. [106]		86.3	96.2	–	–
Zhang et al. [111]		88.8	–	95.9	88.6
Chen et al. [112]		89.6	95.35	90.82	78.51
Jiang et al. [4]		87.1	–	95.8	87.7
Yang et al. [113]		89.1	**97.2**	–	88.7
Bayoumi et al. [114]		**90.2**	**97.2**	**98.9**	**92.8**

The bold values indicate the best R1 accuracy results

computer vision applications such as image classification [113], object re-identification [121], and vehicle re-identification [122]. Wide different vision applications are summarized in a comprehensive survey by Khan et al. [123]. Due to their great modeling capability, they are also introduced for person re-identification either for image-based ReID [124, 125] or for video-based ReID as in [126].

Sarker et al. [127] surveyed the transformer-based ReID systems. They pointed to the challenges that face transformer-based ReID systems. Although it was published in 2024, they mostly reviewed the image-based ReID and only one video ReID system [124] was reviewed through their survey.

Liu et al. [126] used transformers to exploit richer information representation for the raw video by proposing a trigeminal transformer. Three different representations are learned in spatial, temporal, and spatial–temporal views. For each view, local features are extracted. Then, three independent self-view transformers are introduced, and the resulting features are used as input to produce a robust relationship for each view. A cross-view transformer is then introduced to exploit the relationship through different views and aggregate them to get the final video representation. Experiments on iLIDS-VID, PRID2011, and MARS show 91.3%, 96.4%, and 91.2% R1 accuracy, respectively.

Yang et al. [128] introduced the transformer for linking the temporal information between frames by using Spatial Transformer Encode (STM) and Temporal Interaction Module (TIM). STM worked on extracting the spatial features by using an encoder instead of CNN in their work, while TIM discovered the positive guided clues among frames. Further fine-grained module with transformer-based is introduced. Over Mars, 89.2% R1 accuracy is achieved.

Not only the spatiotemporal features could be extracted using transformers; Tang et al. [129] proposed the attribute, identity, and attribute–identity proxy as representative features that are handled using attribute-aware Proxy Embedding Module, and identity-aware Proxy Embedding Module. Results on MARS, DukeMTMC-VideoReID, and iLIDS-VID show 91.8%, 97.4%, and 93.3% R1 accuracy results, respectively.

Transformers need a large dataset for training in order to avoid overfitting problems. To avoid overfitting with a limited dataset, two-stage spatiotemporal transformer with constrained attention is proposed by Zhang et al. [130]. The constrained attention is applied to both spatial and temporal transformers and then an extra global attention

module is added to improve the accuracy. Experiments on MARS, DukeMTMC-VideoReID, and iLIDS-VID show 88.7%, 97.6%, and 87.5% R1 accuracy, respectively.

Zang et al. [131] introduced the transformer as global and local scale feature extractor for the person tiny patches. Each person is divided into small patches then each patch is divided with vertical and horizontal direction to get tiny directive representation. Experiments over Mars and iLIDS-VID show 90.22%, and 92.07% R1 accuracy results.

Self-attention is the key element in transformer, and He et al. [132] used it to learn spatial and temporal features. The vanilla transformer encoder learns some features from each frame, while the vanilla transformer decoder learns the temporal features. The composition of the encoder and decoder forms the proposed dense attention model. They added two extra steps to outperform other's previous work: positional input embedding, and fine-graining step. Positional input embedding investigates the input spatiotemporal features. While fine-graining step works on selecting informative frame features. Over MARS, DukeMTMC-VideoReID, and iLIDS-VID; the proposed work achieved 90.2%, 97.6%, and 92% R1 accuracy, respectively.

In Alsehaim and Breckon [133] introduced temporal clip shift and shuffle module with the global feature learning to get informative frame features. The frames are inducted into this module to form a clip from certain frames, and then they are shuffled. The next video patch part features determine the temporal informative part for each shuffled clip. Over MARS, PRID2011, and iLIDS-VID, the proposed work achieved 96.36%, 96.63%, and 94.67% R1 accuracy, respectively.

Yang et al. [134] used the CNN backbone as a frame-level feature extractor, and then, a mathematical model is proposed to map the continuous frame information into discrete space that is easily discriminated. Then, the sub-sequential filtering module passes only the representative information. Over MARS and DukeMTMC-VideoReID, the proposed work achieved 95.5% and 97.8% R1 accuracy, respectively.

The transformer-based approaches are summarized in Table 7.

3.2.3 Long short-term memory (LSTM)-based approaches

The LSTM is used widely in capturing the temporal information for the video sequence. Courtney et al. [135] used different LSTM networks to select the most appropriate spatial temporal scale for a dataset.

Table 7 The transformer-based approaches (R1 accuracy in %)

Author	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Liu et al. [126]	Transformers-based	91.2	–	**96.4**	91.3
Yang et al. [128]		89.2	–	–	–
Tang et al. [129]		91.8	97.4	–	93.3
Zhang et al. [130]		88.7	97.6	–	87.5
Zang et al. [131]		90.22	–	–	92.07
He et al. [132]		90.2	97.6	–	92
Alsehaim et al. [133]		**96.36**	96.63		**94.67**
Yang et al. in [134]		95.5	**97.8**	–	–

The bold values indicate the best R1 accuracy results

Recently, LSTM is used for cross-view gait information as in [136]. Wu et al. [136] used the LSTM encoder–decoder framework to get the temporal information of the spatial information extracted by CNN. Results over iLIDS-VID and PRID2011 datasets have 41.6% and 69.0% R1 accuracies.

Li et al. [137] proposed a comprehensive model with both LSTM and residual attention components for cross-view gait recognition and measured the accuracy on CASIA gait recognition dataset.

Ouyang et al. [138] used the LSTM in capturing the temporal features. While the spatial features are captured by Two Fusion Streams CNN (TFS-CNN) information. The fusion of TFS-CNN and LSTM produced representative features. Results on iLIDS-VID and PRID2011 have 64.8% and 78.3% R1 accuracy results.

Avola et al. [139] used LSTM to generate person identifying pattern and two dense layers to generate meta-data for the skeleton gait and bone portion of the RGB video stream. Results over iLIDS-VID, PRID2011, and MARS show 73.4%, 82.7%, and 86.5% accuracy, respectively.

Song et al. [140] proposed Extended Global Local Representation learning Network (E-GLRN). E-GLRN mainly depends on Bi-LSTM to extract the temporal local features between consecutive frames in addition to extracting the most informative three frames. Bi-LSTM consisted of forward and backward LSTMs. Bi-LSTM is preferred over the traditional LSTM that uses the previous, current, and future frames. The global and local features are fused together to get a representation for the person. Experiments over iLIDS-VID, PRID2011, and MARS show 81.3%, 91.6%, and 83.3% accuracy, respectively.

Bi-LSTM is also introduced by Dai et al. [141] to get both spatial and temporal features in two different network streams. The first stream with Bi-LSTM and temporal pooling is used to extract the generic features, while the second stream is utilized to extract temporal features of the consecutive frames. Dai, Ju, et al. obtain 80.5% R1 accuracy over MARS, 87.8% over PRID2011, and 57.7% over iLIDS-VID.

Limcharoen [142] proposed the Bi-LSTM for the gait-based dataset and has a competing result on. The LSTM-based approaches are summarized in Table 8.

3.2.4 3D CNN-based approaches

3D CNN is used as position encoding for the temporal information in videos. Bhuiyan et al. [143] introduced Spatial Temporal Cross-Attention (STCA), and 3D CNN is used to get the temporal information of the input tracklets, while the 2D CNN worked as a spatial appearance extractor. A cross-attention module is used further to obtain the

Table 8 The LSTM-based approaches (R1 accuracy in %)

Authors	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Wu et al. [136]	LSTM-based approaches	–	–	69.0	41.6
Ouyang et al. [138]		–	–	78.3	64.8
Avola et al. [139]		**86.5**	–	82.7	73.4
Song et al. [140]		83.3		**91.6**	**81.3**
Dai et al. [141]		80.5	–	–	57.7

The bold values indicate the best R1 accuracy results

salient cues for the fused (2D and 3D) inputs. Comparable results for MARS and iLIDS-VID with 90.3% and 88.3% R1 accuracy results are accomplished.

Gu et al. [47] introduced the 3D CNN as temporal encoder after the 2D appearance extraction CNN. They proposed an appearance-preserving 3D CNN (AP3D) module as a replacement of any 3D convolution network. AP3D shows outperforming R1 accuracy for MARs, and DukeMTMC-VideoReID: 90.7%, and 97.2%, respectively.

Xing et al. [144] used the 3D CNN to recognize the gait in cross-view.

Li et al. [25] introduced compact and multi-scale 3D convolution (M3D) networks that obtain the temporal information. M3D was introduced in a form that it had one spatial kernel and n parallel temporal kernels with variable temporal range. M3D was comprised with residual attention layer (RAL) that refined the resulting temporal cues. M3D is followed by a 2D CNN to get the appearance spatial features. Results on MARs, iLIDS-VID, and PRID2011 show 84.39%, 74%, and 94.4% R1 accuracies.

4 Deep metric learning

Metric learning is an essential component in person re-identification. It aims to calculate the similarity distance between the probe and the gallery extracted features. It is required to minimize the inter-classes variance for the same person, and intra-class variance between different persons. Thus different distance metric calculation is needed.

Zou et al. [145] was the first to review the metric learning algorithms. They categorized them into metric and metric learning methods. The metric methods refer to the calculation of similarity distance between the probe and gallery extracted features [145]. Metric is calculated either from distance metric (Mahalanobis distance, or asymmetric distance metric) or from similarity metric constructed through (hypergraphs, or matching functions), whereas the metric learning usually means constructing the metric matrix from loss objective function design. Figure 13 summarizes the idea.

Zou et al. [145] categorized the metric learning into classical and deep metric learning methods. Several classical metric learning techniques are widely used for video-based re-identification. Such methods as: Keep it Simple and Straight Metric (KISSME) learning algorithms [146]; Cross-view Quadratic Discriminant Analysis (XQDA) distance metric learning [146]; and Local Fisher Discrimination Analysis (LFDA) metric learning [147]. The deep metric learning techniques work on optimizing the loss function to achieve high ReID accuracy. Several loss functions are introduced on video-based

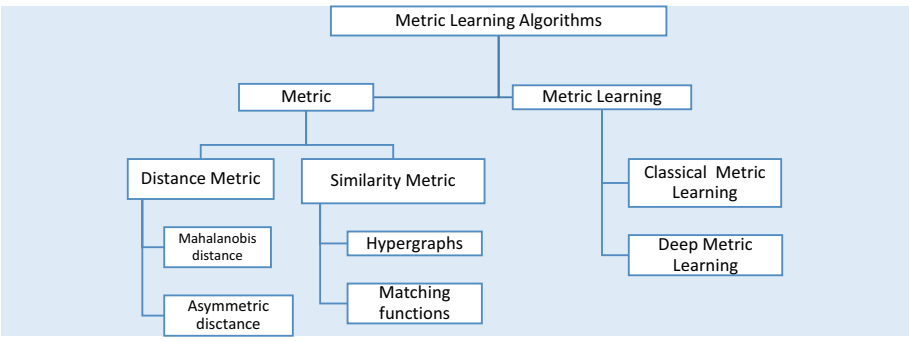

Fig. 13 Metric learning algorithms according to Zou et al. [112] classification

datasets as softmax loss [148], center loss, contrastive loss [149], triplet loss [150], and joint loss. The joint loss works on using more than one loss function in the ReID system. It is a way for improving the single loss results.

Fang et al. [151] modeled the video clips into sets and then employed the triplet loss between them. They used the set theory to calculate the distance between each video clip set rather than the ordinary distance learning which gets the distance between frames rather than a set of frames (clip). Furthermore, hard positive triplet loss is also presented for the proposed clip set. PRID2011, MARS, and iLIDS-VID have 96.6%, 87.8%, and 88.0% R1 accuracy for the proposed set model.

Wang et al. [152] constructed an adaptive metric learning for maximizing the distance between different classes. The adaptive loss function is composed of similarity and regularization terms. The PRID2011 dataset has 73.3% R1 accuracy for the proposed work rather than 15% for the KISSME method.

Yang et al. [153] used a CNN for frame-level features, and then, the temporal relations are modeled into a graph. Softmax cross-entropy loss and triplet loss are concatenated as a final loss function. R1 accuracy over MARS and DukeMTMC-Video ReID is 83.7% and 97.29%, respectively.

A combination of softmax loss function and center loss was introduced by Zhu et al. [154]. The combination of the two losses increased the discriminative capability between the intra- and inter-class variation for the proposed CNN architecture features. Over MARS, 59.8% R1 accuracy is achieved.

Modifications on the conventional triplet loss are introduced by Wojke et al. [155], and they proposed the cosine loss as new metric learning. The cosine loss is created based on softmax loss after simple re-parametrization. Experiments on MARS show 72.93% compared to 71.31% for conventional triplet loss.

Hermans et al. [156] also modified the classical triplet loss by introducing variants. These variants made the hard triplets less important. The hard margin is also replaced by a soft margin. Results on MARS show 81.21% R1 accuracy compared to 79.8% for triplet loss.

Meng et al. [157] discriminated between the person's appearance in the camera view and cross-camera views using Deep Graph Metric Learning (DGML). Two graphs are constructed for spatial and temporal views. Then, the graphs are trained utilizing weak supervision. Results on WL-PRID2011, WL-ILIDs-VID, and WL-MARS are 72.09%, 53.33%, and 68.18% R1 accuracy compared to 51%, 16%, and 21.56% for one-shot learning-based [143], respectively.

5 Deep learning approaches in video-based person ReID

Labeling the unlabeled data is a major problem in deep learning. This section focuses on the published articles for enhancing the learning process for video-based person ReID. Five main approaches are reviewed: supervised learning, unsupervised learning, weakly supervised learning, reinforcement learning, and one-shot learning.

5.1 Supervised video-based person ReID

There are two major obstacles for supervised video labeling: label estimation and feature representation. Despite the existence of labels, most estimation methods are introduced

for image-based ReID. On the other hand, feature representation is a non-trivial task due to frame misalignment, pose change, and illumination variation [21]. Moreover, the supervised labeling needs massive annotation [158]. Supervised video-based approaches are well served through the previous sections.

5.2 Unsupervised video-based person ReID

Unsupervised learning can learn from the unlabeled data, without human guidance, through discovering features and grouping similar ones. Li et al. [159] upgraded their work [160] by proposing an Unsupervised Tracklet Association Learning (UTAL) framework. UTAL is a scalable system that works on the raw data and produces end-to-end labeling for each person without ID duplication or additional preprocessing as in [160]. UTAL starts with per-camera tracklets detection and labels them without applying any verification from the raw video data. Then, extract features from each tracklet. Cross-camera tracklet association CCTA is then performed as a global representation for different views. Experiments over PRID2011, iLIDS-VID, and MARS show: 54.7%, 35.1%, and 49.9% R1 accuracy, respectively (Table 9).

Chen et al. [161] proposed an end-to-end Deep Association Learning (DAL) which is learned from the unlabeled tracklets in two stages: (1) learn the space–time consistency for each tracklet in a single camera where each tracklet is ranked according to the similarity measure. (2) Learn the global consistency across cameras for different tracklets. Experiments over PRID2011, iLIDS-VID, and MARS show: 84.6%, 52.6%, and 46.8% R1 accuracy, respectively.

Ye et al. [162] constructed a Dynamic Graph Matching (DGM) that aimed to estimate labels for each person across different cameras. For each camera, the unlabeled graph is constructed. Each person per frame represents a node. For multiple cameras, it is required to measure the similarity distance between graphs to find the most similar persons (positive samples) and give it an appropriate label. Iterative positive reweighting is performed to enhance the labeling process. Experiments over PRID2011, iLIDS-VID, and MARS show: 89.6%, 55.4%, and 54.3% score estimation accuracy, respectively.

Prasad et al. [163] proposed Spatial Temporal Association Rule-based Deep Annotation free Clustering (STAR-DAC) framework as pure unsupervised labeling framework. STAR works on clustering the visually matched images and then fine-tuning the

Table 9 The deep metric learning publications (R1 accuracy in %)

Author	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Fang et al. [151]	Metric learning	**87.8**	–	**96.6**	**88.0**
Wang et al. [152]		–	–	73.3	–
Yang et al. [153]		83.7	**97.29**	–	–
Zhu et al. [154]		59.8	–	–	–
Wojke et al. [155]		72.93	–	–	–
Hermans et al. [156]		81.21	–	–	–
Meng et al. [157]		WL-MARA 68.18	–	WL-PRID2011 72.09	WL-IIDS-Vid 53.33

The bold values indicate the best R1 accuracy results

clustered one by using some rules. Over DukeMTMC-VideoReID and iLIDS-VID, 66.7% and 59.4% R1 accuracy is achieved, respectively.

On the other hand, Li et al. [164] worked on resolving the problem of assigning wrong pseudo-labels for the same person, which is due to large intra-variation for the common appearance. A multi-granularity pseudo-label prediction method was used at first to predict the label per person image, and then, those labels were used to train expert models that refine the predicted pseudo labels. 81.6%, and 87.3% R1 accuracy for MARS and DukeMTMC-Video ReID is achieved.

Zhang et al. [165] used the camera-aware method for unsupervised learning. CNN was used to extract the unlabeled dataset features, and then, pseudo-labels were assigned for each cluster. The online camera-aware weighting was calculated and updated for only the proxy camera. The selection of individual cameras is recommended in this work rather than all cameras due to complexity issues. This work was tested on Market1501 dataset with 94.1% R1 accuracy.

Kim et al. [166] introduced a novel Disentanglement Switching and Aggregation Network (DSANet). DSANet worked on separating the identity of different cameras regarding the appearance and background similarities. DSANet used frame weight generation based on global pooling to predict the pseudo-labels for each discriminative frame. 91.1% and 97.2% R1 accuracies for Mars and Duke are achieved.

Yang et al. [167] focused on the person's motion in groups in addition to the spatiotemporal features over frames. An Accumulative Motion Context network (AMOC) is developed to get the motion flow from adjacent frames that mainly localize the split person slices in frame sequence first within the camera and then across multiple cameras. The localized slices are trained to utilize an unsupervised manner due to the large number of parameters across cameras. Results on iLIDS-VID, MARS, and DukeMTMC-VideoReID show 52.5%, 65.6%, and 76.8% R1 accuracies, respectively.

To enhance the performance of the unsupervised learning approach, Zeng et al. [168] tried to learn the association between the image and each camera and then update the anchor across cameras to find the positive pairs tracklets. Results on Mars show 73.2% R1 accuracy and 87.0% on DukeMTMC-VideoReID.

Lin et al. [169] highlighted the need for unsupervised learning for video-based person ReID. They surveyed the existing unsupervised approaches from four main challenging viewpoints and the proposed solution perspectives. The four main challenges are related to ground truth unavailability, pseudo-supervision for feature learning, camera invariant-related problems, and dataset gap. Solutions for image-based and video-based ReID enhancement are stated.

Xie et al. [170] worked on learning the video discriminating features from certain informative frames. They used sampling and reweighting strategies for trimming the noisy frames and enhancing the learning accuracy. Frames that include pose changes and partial occlusion, hard frames, are used to improve the clustering accuracy. Experiments over PRID2011, DukeMTMC-VideoReID, and MARS show: 72%, 83%, and 62.7% R1 accuracy, respectively.

Lin et al. [171] followed the iterative clustering and classification methodology for learning the cluster's embedded features. Iterative training (soften similarity learning) is used for clustering hard quantization loss. The unsupervised network is initialized

by setting each image as a separate class, then each group of similar reliable images is grouped, and finally soften labels are assigned to each group and unreliable images are discarded. Experiments over DukeMTMC-VideoReID and MARS show: 76.4% and 62.8% R1 accuracy, respectively.

Wang et al. [172] exploited the different camera network imaging for improving the R1 results. The first step is to determine similar samples delivered by each camera. In the next step, cross-camera views are explored to find the matched pairs between cameras. Experiments over DukeMTMC-VideoReID and MARS show: 76.5% and 65.3% R1 accuracy, respectively.

Xie et al. [173] proposed two sequential steps to improve the ReID ranking results. The first step is removing the noisy frames that include high occlusion using a dynamic threshold for each dataset. In the second step, hard frames that include pose changes are used to train the model. Experiments over DukeMTMC-VideoReID and MARS show: 82.8% and 61.8% R1 accuracy, respectively.

5.3 Weakly supervised video-based person ReID

Weakly supervised learning is recently used for image-based ReID [27, 174], localization [175], and for video-based ReID [176–178]. Weakly supervised learning means identifying the person's identity in the video regardless of the dedicated annotation in each video frame. Each set of probe video clips has several persons which is called a bag. The goal is to label the bag instead of labeling the individual frames. Therefore, it is considered a multi-instance multi-labeling learning problem [177]

Wang et al. [176] used the weakly supervised technique to label a person within a bag with his corresponding identity. For this purpose, they constructed two datasets: WL-MARS and WL-DUKE, from the original MARS and DukeMTMC-VideoReID datasets, employing aggregating more than one person in one tracklet per video. They have built their model to label each person in the given video through: CNN and two sequential steps (coarse and fine-grained). The coarse-grained aimed at labeling identities from certain tracklets in the video. However, the fine-grained identified the person from the tracklet. The results over WL-Mars and WL-DUKE are 65.0% and 70.2% R-1 accuracy, respectively.

Meng et al. [177] proposed a cross-view of multi-person/label in multiple camera views (CV-MIMl) such that they could find the person intra-bag (with the same bag) and cross other bags of different camera views utilizing clustering. The intra-bag is initially constructed to determine whether the same person has more than one image in the bag or not. The cross-view afterward clusters all person images across different camera views. Mask R-CNN [179] is implemented to generate the bounding box for each person with a confidence score. WL-DUKE, WL-MARS, WL-PRID2011, and WL-iLIDS-VID datasets are evaluated over the CV-MIML model and have 78.05%, 66.88%, 72%, and 60% R1 accuracy, respectively.

Yu et al. [178] proposed a feature rectifier for the weakly supervised labels. First, the initial weakly supervised model assigned pseudo-labels for different identities, and then, the features were learned. Due to illumination and appearance changes in frame sequences, the feature learning is distorted. A further decision feature boundary has to

be rectified using the rectification function. The model achieved 72.1% R1 accuracy for the DukeMTMC-VideoReID dataset.

Liu et al. [179] removed noisy tracklets to improve the ReID results. After obtaining bags and assigning a label for each bag, noisy tracklets are removed. Afterward, learning the cross-bag tracklet association is used to differentiate between positive and negative pairs. Experiments over MARS and DukeMTMC-VideoReID have 88.1% and 90.2% R1 accuracy results, respectively.

5.4 One-shot learning for video-based person ReID

According to the aforementioned reasons, Liu et al. [180] used one-shot learning as an initial labeling model to build a labeling system. The initial model learned the discriminative features that could generate the pseudo-labels for the unlabeled tracklets. Afterward, dynamic high-confidence samples are selected to update the initial model. Learning of one-shot pseudo-labeling is described in [181]. The experiments were held over DukeMTMC-VideoReID and MARS and attained 89.2% and 66.7% R1 accuracy, respectively.

5.5 Reinforcement learning for video-based person ReID

Reinforcement learning (RL) is introduced in image-based ReID by Lan et al. [103]. Then, Zhang et al. [182] introduced it for the video-based ReID. Zhang et al. [182] used the RL rewarding rules to instantly handle only two images at a time to decide whether they are similar or not. The results show 71.2%, 85.2%, and 60.2% R1 accuracy for MARS, PRID2011, and iLIDS-VID datasets, respectively.

Ouyang et al. [31] introduced a Self-Paced Learning (SPL) model [183] to select the appropriate frames. The SPL model is introduced by Kumar [183] to optimize the curriculum learning process. SPL depends on learning from the simple examples to the hardest ones. Behind this concept, Ouyang et al. [31] extracted the spatial temporal information employing a network composed of CNN and LSTM and then a SPL module was used further to update the network parameters to make it as mature and stable as possible. Results on MARS, PRID2011, and iLIDS-VID show 74.8%, 85.3%, and 70.5% R1 accuracy, respectively.

Zhang et al. [184] noticed that there are similar repetitive and redundant frames on some datasets such as MARS. Those frames represent noisy information as similar scenes appear. For that purpose, reinforcement learning (RL) is introduced to find the most relevant frame. The agent is trained to find the appropriate one. The results show 83%, 91.2%, and 68.4% R1 accuracy for MARS, PRID2011, and iLIDS-VID datasets, respectively.

6 Discussion

This section highlights some challenges that affect the video-based ReID performance according to the reviewed publications. The discussion is presented from two viewpoints: The first focuses on the video-based person ReID systems architectures, strengths, and weakness points, whereas the second viewpoint addresses the performance evaluation over the four benchmark datasets.

6.1 General video-based ReID system discussion

Since the video-based ReID system performance is affected by the extracted spatial and temporal feature types, both are investigated in this section.

Different approaches are proposed to provide representative spatial and temporal information for each tracklet. The spatial features have many forms as appearance-based, attribute-based, and motion-based features. Some of these features are extracted through some steps before they are introduced to the ReID system. For example, attribute-based features require extra annotation to the dataset according to the used number of attributes and their type (stationary or dynamic). On the other hand, the pose features require estimating the person's key points and the reference pose, while the gait-based features require getting the RGB image silhouettes before being merged into the ReID system. Although the appearance features are demonstrative, it is preferred to be merged with additional motion-relevant features such as optical flow or gait to obtain better results. Although the spatial features express representative information of frames, the temporal features act as a detailed connective link across individual frames. The temporal features represent redundant and repetitive information which has two side effects for video-based ReID implementation: (1) They add extra computational cost; (2) they provide huge information due to redundancy that may disturb the system. Thus, further techniques are introduced to handle these two points. Redundancy reduction is performed by selecting the most salient and representative frames either by reweighting frames or by removing the noisy ones. Thus, concentrating only on selecting the salient frames reduces the computational cost.

On the other side, other researchers develop architectures to enhance the video-based ReID capability of acquiring the information details using attention models or transformers. Both approaches show promising results over different datasets [134]. Uses discretizing the frames to select the most relevant frames then transforms are utilized to learn frames sequencing. As a result, outstanding state-of-the-art accuracy is obtained.

As shown previously, transformers have an informative representation of video-based ReID. It obtains the highest R1 accuracy results for the challenging MARS dataset. However, as transformers are initially presented for sequence representation in NLP and recently utilized in video sequencing in addition to its consuming computation cost, it was not widely addressed in video-based person ReID.

Besides, 3D CNN is used as a temporal feature aggregator. It acted as an encoder for the temporal information instead of just extracting the spatial features as in traditional 2D CNN. Krichen et al. [185] argued that 3D CNN outperforms CNN and LSTM for temporal encoding. In some cases, temporal features face unavoidable appearance misalignment for consecutive frames. Thus, techniques such as 3D CNN and graph CNN are appropriate to reduce this misalignment.

6.2 Benchmark datasets evaluation over state-of-the-art approaches

This subsection declares the performance of the various video-based person ReID architectures for each benchmark dataset. According to the deep learning-based categories: supervised-based learning including (LSTM, 3D CNN, attention models, and transformers), unsupervised learning, weekly supervised learning, and one-shot learning approaches, a detailed comparison is held. For each deep learning category, the highest

Table 10 The unsupervised learning publications (R1 accuracy %)

Author	Method	MARS	DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Li et al. [159]	Unsupervised learning	49.9	–	54.7	35.1
Chen et al. [161]		46.8	–	84.6	52.6
Ye et al. [162]		54.3	–	**89.6**	55.4
Prasad et al. [163]		–	66.7	–	**59.4**
Yang et al. [167]		65.6	76.8	–	52.5
Zeng et al. [168]		**73.2**	**87.0**	–	–
Xie et al. [170]		62.7	83	72	–
Lin et al. [171]		62.8	76.4	–	–
Wang et al. [172]		65.3	76.5	–	–
Xie et al. [173]		61.8	82.8	–	–

The bold values indicate the best R1 accuracy results

Table 11 The weakly supervised learning publications (R1 accuracy in %)

Author	Method	WI-MARS	WI-DukeMTMC-VideoReID	PRID2011	iLIDS-VID
Wang [163]	Weakly supervised learning	65.0	70.2	–	–
Meng et al. [177]		66.88	78.05	**72**	**60**
Yu et al. [165]		–	72.1	–	–
Liu et al. [179]		**88.1**	**90.2**	–	–

The bold values indicate the best R1 accuracy results

reviewed R1 accuracy is reported. Then, detailed comparison between these categories is accomplished over four benchmark datasets regarding the highest R1. Despite of achieving higher results for supervised learning than other learning-based categories, the unsupervised learning approaches could not be neglected due to its learning capability from unlabeled dataset. The unsupervised learning is a suitable learning methodology for unlabeled datasets. Thus, it is important to highlight the gap between the supervised and unsupervised learning approaches by analyzing the reviewed publications over the four benchmark datasets.

6.2.1 MARS dataset

MARS is one of the most challenging datasets due to the variety of poses, occlusion, scale, lightning, background change, and bounding box misalignment. These challenges causes network inconsistency. Table 12 shows the highest R1 results per learning approach over MARS. For transformers, Table 7 shows that the highest R1 accuracy results achieved by Alsehaim et al. are 96.36% [133]. For attention, as shown in Table 6, the highest R1 accuracy result achieved by Bayoumi et al. is 90.2% [114]. For LSTM, Table 8 shows the highest R1 accuracy results achieved by Avola et al. are 86.5% [139]. Using 3D CNN (by Gu et al. [47]) obtained the highest R1 accuracy that reached 90.7% [47]. On the other hand, with the massive work for improving the unsupervised learning approach results, Table 10 shows that Zeng et al. [168] outperform other next-best unsupervised approaches and have 73.1%. The weakly supervised (summarized in Table 11) reached the highest R1 accuracy of 88.1%. Liu et al. [179] has 88.1%, while the one-shot

Table 12 MARS state-of-the-art R1 accuracy (%)

Year	Author	Deep learning approach	R1 accuracy (%)
2023	Alsehaim et al. [133]	Transformer	**96.36**
2022	Bayoumi et al. [114]	Attention	90.2
2022	Zeng et al. [168]	Unsupervised	73.1
2022	Gu et al. [47]	3D CNN	90.7
2023	Liu et al. [179]	Weakly supervised	88.1
2023	Kim et al.[21]	One shot	66.7
2020	Avola et al. [139]	LSTM	86.5

The bold value indicates the best R1 accuracy result

learning reported by Kim et al. [21] has 66.7%. So, these results show that the transformer proposed by [134] has superiority over other deep-learning architectures for MARS dataset (Table 12). Figure 14 summarizes the results.

6.2.1.1 DukeMTMC-VideoReID dataset DukeMTMC-VideoReID is less addressed by research works compared to MARS. Table 13 shows the highest R1 results per learning approach over DukeMTMC-VideoReID. For transformers, Table 7 shows that the highest R1 accuracy result is achieved by Yang et al. (97.8%) [134]. For attention, as shown in Table 6, the highest R1 accuracy result is achieved by Yang et al. [113], and Bayoumi et al. (97.2%) [114]. Using 3D CNN (by Gu et al. [47]) obtained the highest R1

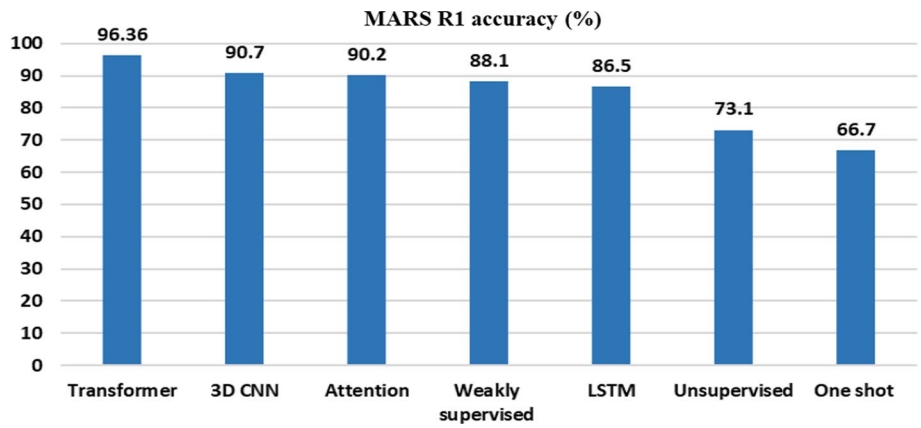

Fig. 14 Different deep video-based person ReID performance evaluation on MARS

Table 13 DukeMTMC-VideoReID state of the art R1 accuracy (%)

Year	Author	Deep learning approach	R1 accuracy (%)
2024	Yang et al. [134]	Transformer	**97.8**
2020	Yang et al. [153]	Unsupervised	97.29
2022	Yang et al. [113], Bayoumi et al. [114]	Attention	97.2
2020	Gu et al. [47]	3D CNN	97.2
2023	Liu et al. [179]	Weakly supervised	90.2
2022	Liu et al. [168]	One shot	89.2
–	–	LSTM	–

The bold value indicates the best R1 accuracy result

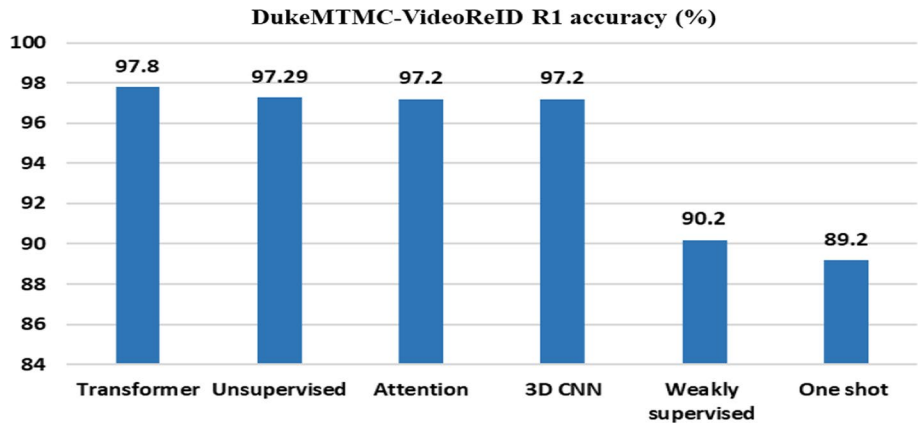

Fig. 15 Different deep video-based person ReID performance evaluations on DukeMTMC-VideoReID

Table 14 PRID2011 state-of-the-art R1 accuracy (%)

Year	Author	Deep learning approach	R1 results accuracy (%)
2021	Liu et al. [126]	Transformer	96.4
2022	Bayoumi et al. [114]	Attention	**98.9**
2019	Ye et al. [162]	Unsupervised	89.6
2020	Li et al. [25]	3D CNN	94.4
2019	Meng et al. [177]	Weakly supervised	72
–	–	One shot	–
2019	Song et al. [140]	LSTM	91.6

The bold value indicates the best R1 accuracy result

accuracy that reached 97.2%. On the other hand, Table 10 shows that the best unsupervised learning result is achieved by Yang et al. [153] (97.29%), while weakly supervised learning (summarized in Table 11) reached the highest R1 accuracy result of 90.2% achieved by Liu et al. [179]. One-shot learning, reported by Kim et al. [21], has 89.2% accuracy. Thus, the superiority results show that the transformer in [134] has outperformed the results of DukeMTMC-VideoReID. Figure 15 summarizes the results.

6.2.1.2 PRID2011 dataset Table 14 summarizes the PRID2011 state-of-the-art video-based ReID systems. For transformers, Table 7 shows that PRID2011 is only tested by Liu et al. and has 96.4% [126]. For attention, as shown in Table 6, the highest R1 accuracy result achieved by Bayoumi et al. is 98.9%. For LSTM, Table 8 shows the highest R1 accuracy results achieved by Song et al. [140] are 91.6%. For 3D CNN, 3D CNN was only tested by Li et al. [25] and obtained 94.4%. On the other hand, Table 10 shows the best unsupervised learning results achieved by Ye et al. [162] are 89.6, while weakly-supervised learning (summarized in Table 11) shows that PRID2011 is only tested by Meng et al. [177] and obtained 72%. The one-shot learning is not tested over PRID2011. So attention proposed in [114] has superiority results on PRID2011. Figure 16 shows the results.

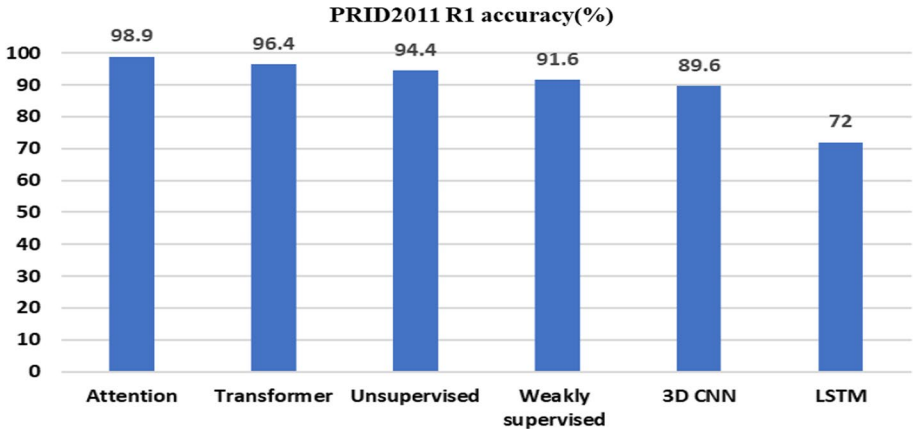

Fig. 16 Different deep video-based person ReID performance evaluations on PRID2011

Table 15 ILIDs-VID state-of-the-art R1 accuracy (%)

Year	Author	Deep learning approach	R1 results accuracy (%)
2023	Alsehaim et al. [133]	Transformer	**94.67**
2022	Bayoumi et al. [114]	Attention	92.8
2022	Prasad et al. [163]	Unsupervised	59.4
2022	Bhuiyan et al. [143]	3D CNN	88.3
2019	Meng et al. [177]	Weakly supervised	60
–	–	One shot	–
2019	Song et al. [140]	LSTM	81.3

The bold value indicates the best R1 accuracy result

6.2.1.3 ILIDS-VID dataset ILIDS-VID has been tested in many publications. Table 15 summarizes the state-of-the-art video-based ReID systems on iLIDS-VID. For transformers, Table 7 shows that the highest R1 accuracy results achieved by Alsehaim, Aishah et al. are 94.67% [133]. For attention, as shown in Table 6 the highest R1 accuracy results achieved by Bayoumi et al. are 92.8%. For LSTM, Table 8 shows the highest R1 accuracy results achieved by Song et al. [140] are 81.3%. For 3D CNN, Bhuiyan et al. [143] obtained the highest R1 accuracy results by 88.3%. On the other hand, the best unsupervised learning results achieved by Prasad et al. [163] are 59.6%, as summarized in Table 10, while weakly supervised results in Table 11 show that Meng et al. in [177] are the only one who tested iLIDS-VID and have 60%. The one-shot learning is not tested over iLIDS-VID. So transformer proposed in [133] has superior results for iLIDS-VID. Figure 17 shows the results of iLIDS-VID.

From the previous analysis, it is noticed that MARS needs massive work for unsupervised, weakly supervised, and one-shot learning, while DukeMTMC-VideoReID needs more enhancements for weakly-supervised and one-shot learning. On the other hand, more improvements can be performed on PRID2011 using LSTM and 3D CNN, whereas more enhancements for LSTM, unsupervised and weakly supervised, can result in promising accuracies for iLIDS-VID. Since transformers have competing

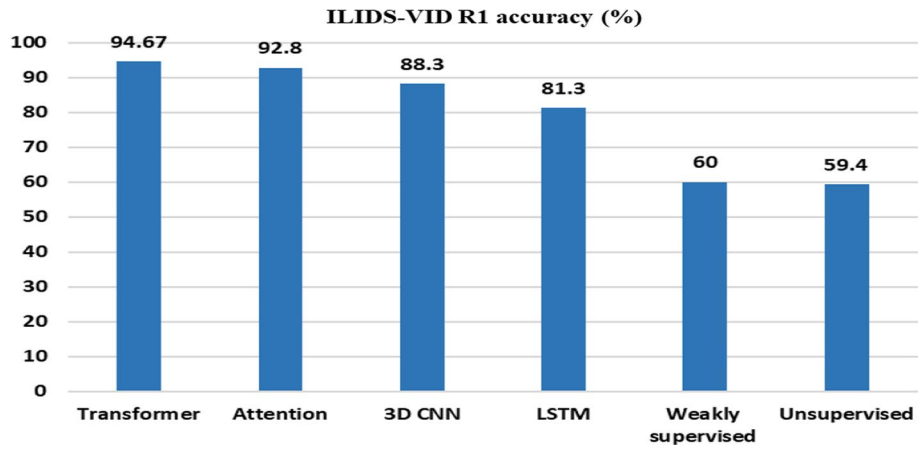

Fig. 17 Different deep video-based person ReID performance evaluation on iLIDS-VID

results on the four benchmark datasets, there is a need to be more served in video-based person ReID to get better R1 results. Also, extra experiments are needed to reduce its computational cost.

Attention models have competing results also, so modifications over attention models are welcomed to have better results.

7 Challenges and future directions

Person ReID is still considered as an open challenging problem. As recommended by [186], it promotes the research toward other research areas such as player re-identification, crowd management, criminal search, and civilian and military camouflage clothing ReID [186]. Video Person ReID has some open challenging research areas as:

7.1 Cloth changing problem

One of the most challenging problems is the long-term cloth-changing which means identifying a person despite appearance variations due to cloth-changing. To discuss this issue, it is required to record videos for days and months under the probability of a person's clothes changing under inconsistency issue as discussed in [187, 188]. Thus, constructing a new realistic dataset as Real28 and Long-Term Cloth Changing (LTCC) datasets in [187, 188], respectively, is presented. Another challenging problem is uniform discrimination, i.e., extracting a person of interest having a dress code similar to the other persons. Similar appearance features for all persons are presented. These two problems require more investigation to get reasonable retrieval results.

7.2 Uniform discrimination

Another challenging problem is uniform discrimination, i.e., extracting a person of interest through those having similar dress code. Similar appearance features for all persons are presented. This problem has been addressed in schools, factories, and companies with the same appearance. Extra experiments are needed in this direction and results can be evaluated on FGPR dataset [11].

7.3 Existing systems performance improvements

As mentioned in the discussion section, there are many improvements for each approach group that are recommended to achieve better results. In addition, it is recommended to examine using the 3D modeling for person representation and apply different temporal feature backbones to the ReID systems over the four benchmark datasets. The results obtained in [82] show promising results.

In the era of augmented data, generative adversarial networks, GAN [185] (used for generating datasets such as [189] and [190]), it is recommended to build more generic ReID systems that achieve high accuracy on videos generated by those models. GAN is also presented to help in ReID model training as Jot-GAN [191]. A recommendation to apply such techniques is welcomed for video-based approaches.

For transformer computational cost reduction, Chen et al. [192] succeeded to reduce the computational cost by introducing ResNet-50 as a former head part to the transformer but it is applied for image-based ReID. Also, Wang et al. [193] and Zhou et al. [194] used the transformer of the image-based ReID targeting finer representation. Accordingly, these approaches are highly recommended for video-based ReID.

7.4 Datasets resources and models

There is a need to increase the number of publicly available video datasets and support annotation. On the other hand, it is essential to the build models that are more generic and covers different situations such as open and closed environment situations with an adequate number of high-resolution cameras for a large number of persons. As a result, a scalable video ReID that could process these large datasets is mandatory. A hybrid architecture of transformers, attention, and 3D CNN would help in building these models.

8 Conclusion

Video-based person re-identification plays a key role in video surveillance applications, tracking, finding lost people, and criminal investigation. Although deep learning techniques have been involved in building video-based person ReID systems, they are less served compared to image-based techniques. Thus, up to our knowledge, this is the first survey that focuses only on deep video-based person ReID according to ReID system implementation workflow. The survey reviews the video-based person ReID datasets (benchmark and special purpose), deep feature learning (spatial and temporal), deep metric learning, and the deep-learning backbone architecture approaches (supervised, unsupervised, weakly supervised, and one-shot learning). The proposed survey provides recommendations to improve the outcomes depending on the various strengths found in each technique. Over four benchmark datasets, a comparative analysis for the state-of-the-art approaches is performed. In addition, some future research directions are suggested according to the analyzed approaches performance. The analysis shows that more developments are needed for unsupervised, weakly supervised, and one-shot, deep learning approaches to achieve higher results. And transformers achieved superior results on the four benchmark datasets.

Acknowledgements

Its purpose is to thank Electronics Research Institute technical support team who helped with the research but did not qualify for authorship.

Author contributions
All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by RSMS, MMM and NSA, HF, and SM.

Funding
Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 8 December 2023 Accepted: 11 March 2024
Published online: 15 May 2024

References

1. M.O. Almasawa, L.A. Elrefaei, K. Moria, A survey on deep learning-based person re-identification systems. IEEE Access **7**, 175228–175247 (2019)
2. M. Ye et al., Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 2872–2893 (2021)
3. P. Dedeepya, Recent trends in person re-identification: an overview. Turk. J. Comput. Math. Edu. (TURCOMAT) **12**(9), 1841–1846 (2021)
4. X. Jiang et al., Rethinking temporal fusion for video-based person re-identification on semantic and time aspect. Proc. AAAI Conf. Artif. Intell. **34**(07), 11133–11140 (2020)
5. H. Wang et al., A comprehensive overview of person re-identification approaches. IEEE Access **8**, 45556–45583 (2020)
6. Q. Leng, M. Ye, Q. Tian, A survey of open-world person re-identification. IEEE Trans. Circuits Syst. Video Technol. **30**(4), 1092–1108 (2019)
7. E. Yaghoubi, A. Kumar, H. Proença, SSS-PR: A short survey of surveys in person re-identification. Pattern Recogn. Lett. **143**, 50–57 (2021)
8. A. Yadav, D.K. Vishwakarma, Person re-identification using deep learning networks: a systematic review. arXiv preprint https://arxiv.org/abs/2012.13318 (2020)
9. D. Wu et al., Deep learning-based methods for person re-identification: a comprehensive review. Neurocomputing **337**, 354–371 (2019)
10. L. Jiao et al., New generation deep learning for video object detection: a survey. IEEE Trans. Neural Netw. Learn. Syst. **33**, 3195–3215 (2021)
11. J. Yin, A. Wu, W.-S. Zheng, Fine-grained person re-identification. Int. J. Comput. Vis. **128**(6), 1654–1672 (2020)
12. A. Doering et al., PoseTrackReID: dataset description. arXiv preprint http://arxiv.org/abs/2011.06243 (2020)
13. L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, MARS: a video benchmark for large-scale person re-identification, in *ECCV* (2016)
14. A. Dehghan, S.M. Assari, M. Shah, Gmmcp tracker: globally optimal generalized maximum multi clique problem for multiple object tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015)
15. http://zhenglab.cecs.anu.edu.au/Project/project_mars.html
16. M. Hirzer et al., Person re-identification by descriptive and discriminative classification, in *Scandinavian Conference on Image Analysis* (Springer, 2011)
17. https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/prid11/
18. https://xiatian-zhu.github.io/downloads_qmul_iLIDS-VID_ReID_dataset.html.
19. T. Wang et al., Person re-identification by video ranking, in *European Conference on Computer Vision* (Springer, Cham, 2014)
20. https://exposing.ai/duke_mtmc/
21. Y. Wu et al., Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
22. P. Xu, X. Zhu, Long-term person re-identification: a benchmark. arXiv e-prints (2021), arXiv-2105
23. S. Yu et al., Cocas: a large-scale clothes changing person dataset for re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
24. https://github.com/PengBoXiangShang/deepchange
25. J. Li, S. Zhang, T. Huang, Multi-scale temporal cues learning for video person re-identification. IEEE Trans. Image Process. **29**, 4461–4473 (2020)
26. M. Andriluka et al., Posetrack: a benchmark for human pose estimation and tracking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
27. G. Wang et al., Weakly supervised person re-ID: differentiable graphical learning and a new benchmark. IEEE Trans. Neural Netw. Learn. Syst. **32**(5), 2142–2156 (2020)

28. S. Zhang et al., RealGait: gait recognition for person re-identification. arXiv preprint http://arxiv.org/abs/2201.04806 (2022)
29. D. Davila et al., MEVID: multi-view extended videos with identities for video person re-identification, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023)
30. K. Corona et al., Meva: a large-scale multiview, multimodal video dataset for activity detection, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021)
31. D. Ouyang, J. Shao, Y. Zhang, Y. Yang, H.T. Shen, Video-based person re-identification via self-paced learning and deep reinforcement learning framework, in *Proceedings of ACM Multimedia Conference (MM)* (2018), pp. 1562–1570
32. Y. Zhao et al., Attribute-driven feature disentangling and temporal aggregation for video person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019)
33. S. Xu et al., Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in *Proceedings of the IEEE International Conference on Computer Vision* (2017)
34. X. Zhang, F. Pala, B. Bhanu, Attributes co-occurrence pattern mining for video-based person re-identification, in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (IEEE, 2017)
35. Z. Yang et al., Incremental XQDA metric learning for person reidentification., in *2018 IEEE International Conference on Information and Automation (ICIA)* (IEEE, 2018)
36. J. Yin, A. Wu, W.-S. Zheng, Fine-grained person re-identification. Int. J. Comput. Vis. **128**, 1654–1672 (2020)
37. W. Song et al., Discriminative feature extraction for video person re-identification via multi-task network. Appl. Intell. **51**(2), 788–803 (2021)
38. S. Liao et al., Person re-identification by local maximal occurrence representation and metric learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015)
39. Z. Chen et al., Attribute-aware identity-hard triplet loss for video-based person re-identification. arXiv preprint http://arxiv.org/abs/2006.07597 (2020)
40. W. Song et al., A two-stage attribute-constraint network for video-based person re-identification. IEEE Access **7**, 8508–8518 (2019)
41. T. Chai et al., Video person re-identification using attribute-enhanced features. arXiv preprint http://arxiv.org/abs/2108.06946 (2021)
42. W. Song et al., Video-based person re-identification using a novel feature extraction and fusion technique. Multimed. Tools Appl. **79**, 12471–12491 (2020)
43. D. Li et al., A richly annotated dataset for pedestrian attribute recognition. arXiv preprint http://arxiv.org/abs/1603.07054 (2016)
44. Z. Chen, A. Li, Y. Wang, A temporal attentive approach for video-based pedestrian attribute recognition, in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (Springer, Cham, 2019)
45. W. Zhang et al., Learning compact appearance representation for video-based person re-identification. IEEE Trans. Circuits Syst. Video Technol. **29**(8), 2442–2452 (2018)
46. T. Wang et al., Person re-identification by video ranking, in *European Conference on Computer Vision* (Springer, Cham, 2014)
47. X. Gu et al., Appearance-preserving 3d convolution for video-based person re-identification, in *European Conference on Computer Vision* (Springer, Cham, 2020)
48. J. Li, S. Zhang, T. Huang, Multi-scale 3d convolution network for video based person re-identification. Proc. AAAI Conf. Artif. Intell. **33**(01), 8618–8628 (2019)
49. J. Li et al., Global-local temporal representations for video person re-identification, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019)
50. X. Liao et al., Video-based person re-identification via 3d convolutional networks and non-local attention, in *Asian Conference on Computer Vision* (Springer, Cham, 2018)
51. Y.-J. Cho, K.-J. Yoon, Improving person re-identification via pose-aware multi-shot matching, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016)
52. M.S. Sarfraz et al., A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
53. A. Bhuiyan et al., Pose guided gated fusion for person re-identification, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020)
54. H. Zhao et al., Spindle net: person re-identification with human body region guided feature decomposition and fusion, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
55. D. Wei et al., Pose-guided multi-scale structural relationship learning for video-based pedestrian re-identification. IEEE Access **9**, 34845–34858 (2021)
56. C. Gao et al., Pose-guided spatiotemporal alignment for video-based person re-identification. Inf. Sci. **527**, 176–190 (2020)
57. Pan, H., et al. "Pose-Aided Video-based Person Re-Identification via Recurrent Graph Convolutional Network." *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
58. C. Su et al., Pose-driven deep convolutional model for person re-identification, in *Proceedings of the IEEE International Conference on Computer Vision* (2017)
59. M. Snower et al., 15 keypoints is all you need, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
60. E. Insafutdinov et al., Deepercut: a deeper, stronger, and faster multi-person pose estimation model, in *European Conference on Computer Vision* (Springer, Cham, 2016)
61. L. Zheng et al., Pose-invariant embedding for deep person re-identification. IEEE Trans. Image Process. **28**(9), 4500–4509 (2019)
62. A. Elaoud et al., Person re-identification from different views based on dynamic linear combination of distances. Multimed. Tools Appl. **80**(12), 17685–17704 (2021)
63. T. Bendokat, R. Zimmermann, P.-A. Absil, A Grassmann manifold handbook: basic geometry and computational aspects. arXiv preprint http://arxiv.org/abs/2011.13699 (2020)

64. X. Hu et al., Hypergraph video pedestrian re-identification based on posture structure relationship and action constraints. Pattern Recogn. **111**, 107688 (2021)
65. Z. Cao et al., Realtime multi-person 2d pose estimation using part affinity fields, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
66. J. Liu et al., Spatial–temporal correlation and topology learning for person re-identification in videos, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021)
67. D. Wei, Z. Wang, Y. Luo, Video person re-identification based on RGB triple pyramid model. Vis. Comput. **39**, 501–517 (2022)
68. Y. Xian et al., Graph-based self-learning for robust person re-identification, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023)
69. J. Lu et al., Exploring high-order spatio-temporal correlations from skeleton for person re-identification. IEEE Trans. Image Process. **32**, 949–963 (2023)
70. A. Nambiar, A. Bernardino, J.C. Nascimento, Gait-based person re-identification: a survey. ACM Comput. Surv. (CSUR) **52**(2), 1–34 (2019)
71. A. Sepas-Moghaddam, A. Etemad, Deep gait recognition: a survey. arXiv preprint http://arxiv.org/abs/2102.09546 (2021)
72. S. Zheng et al., Robust view transformation model for gait recognition, in *2011 18th IEEE International Conference on Image Processing* (IEEE, 2011)
73. H. Iwama et al., The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans. Inf. Forensics Secur. **7**(5), 1511–1521 (2012)
74. N. Takemura et al., Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans. Comput. Vis. Appl. **10**(1), 1–14 (2018)
75. O. Elharrouss et al., Gait recognition for person re-identification. J. Supercomput. **77**(4), 3653–3672 (2021)
76. H. Rao et al., A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 6649–6666 (2021)
77. B. Rahi, M. Li, M. Qi, A review of techniques on gait-based person re-identification. Int. J. Netw. Dyn. Intell. **2**(1), 66–92 (2023)
78. Y. Zhao et al., Gait-assisted video person retrieval. IEEE Trans. Circuits Syst. Video Technol. **33**(2), 897–908 (2022)
79. Z. Chang et al., Seq-masks: bridging the gap between appearance and gait modeling for video-based person re-identification, in *2021 International Conference on Visual Communications and Image Processing (VCIP)* (IEEE, 2021)
80. L. Wang et al., Fusing the appearance and gait features for clothes-changing video person re-identification. Available at SSRN 4718125
81. V.D. Nguyen et al., Attention-based shape and gait representations learning for video-based cloth-changing person re-identification. arXiv preprint http://arxiv.org/abs/2402.03716 (2024)
82. V.D. Nguyen, P. Mantini, S. K. Shah, Temporal 3D Shape Modeling for Video-based Cloth-Changing Person Re-Identification, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024)
83. A. Dosovitskiy et al., FlowNet: learning optical flow with convolutional networks, in *Proceedings IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 2758–2766
84. N. McLaughlin, J.M. Del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016)
85. M. Kiran et al., Flow guided mutual attention for person re-identification. Image Vis. Comput. **113**, 104246 (2021)
86. S. Cho, H. Foroosh, Spatio–temporal fusion networks for action recognition, in *Asian Conference on Computer Vision* (Springer, Cham, 2018)
87. D. Chung, K. Tahboub, E.J. Delp, A two stream Siamese convolutional neural network for person re-identification, in *Proceedings of the IEEE International Conference on Computer Vision* (2017)
88. G. Chen et al., Learning recurrent 3D attention for video-based person re-identification. IEEE Trans. Image Process. **29**, 6963–6976 (2020)
89. D. Chen et al., Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
90. R. Hou et al., IAUnet: global context-aware feature learning for person reidentification. IEEE Trans. Neural Netw. Learn. Syst. **32**, 4460–4474 (2020)
91. A. Vaswani et al., Attention is all you need. Adv.Neural Inf. Process. Syst. 5998–6008 (2017)
92. S. Li et al. Diversity regularized spatiotemporal attention for video-based person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
93. Q. Zhou et al., Attention-based neural architecture search for person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**(11) (2021)
94. X. Song, Z. Jin, Domain adaptive attention-based dropout for one-shot person re-identification. Int. J. Mach. Learn. Cybern. **13**, 1–14 (2021)
95. D. Wu et al., Attention deep model with multi-scale deep supervision for person re-identification. IEEE Transactions on Emerging Topics in Computational Intelligence **5**(1), 70–78 (2021)
96. G. Zhang et al., Hybrid-attention guided network with multiple resolution features for person re-identification. Inf. Sci. **578**, 525–538 (2021)
97. C. Wang et al., Recurrent deep attention network for person re-identification, in *2020 25th International Conference on Pattern Recognition (ICPR)* (IEEE, 2021)
98. S. Chen, H. Zhang, Person re-identification based on frequency channel attention networks under the surveillance scenario. J. Phys. Conf. Ser. **1966**(1), 012025 (2021)
99. C. Wang, G. Zhang, W. Zhou, Deep progressive attention for person re-identification. J. Electron. Imaging **30**(4), 043028 (2021)
100. Q. Zhao et al., Part-level attention networks for cross-domain person re-identification. *IET Image Process.* (2021)
101. X. Lin et al., Diff attention: a novel attention scheme for person re-identification. Comput. Vis. Image Underst. **228**, 103623 (2023)

102. X. Ning et al., JWSAA: joint weak saliency and attention aware for person re-identification. Neurocomputing **453**, 801–811 (2021)
103. X. Lan et al., Deep reinforcement learning attention selection for person re-identification. arXiv preprint http://arxiv.org/abs/1707.02785 (2017)
104. G. Chen et al., Spatial-temporal attention-aware learning for video-based person re-identification. IEEE Trans. Image Process. **28**(9), 4192–4205 (2019)
105. Y. Liu, J. Yan, W. Ouyang, Quality aware network for set to set recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
106. Y. Fu et al., STA: Spatial-temporal attention for large-scale video-based person re-identification. Proc. AAAI Conf. Artif. Intell. **33**(01), 8287–8294 (2019)
107. Z. Zhou et al., See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
108. X. Shu et al., Diverse part attentive network for video-based person re-identification. Pattern Recognit. Lett. **149**, 17–23 (2021)
109. A. Subramaniam, A. Nambiar, A. Mittal, Co-segmentation inspired attention networks for video-based person re-identification, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019)
110. L. Wu et al., Where-and-when to look: Deep Siamese attention networks for video-based person re-identification. IEEE Trans. Multimed. **21**(6), 1412–1424 (2018)
111. Z. Zhang et al., Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
112. C. Chen et al., Learning discriminative features with a dual-constrained guided network for video-based person re-identification. Multimed. Tools Appl. **80**, 28673–28696 (2021)
113. F. Yang et al., Relation-based global-partial feature learning network for video-based person re-identification. Neurocomputing **488**, 424–435 (2022)
114. R.M. Bayoumi et al., Person re-identification via pyramid multipart features and multi-attention framework. Big Data Cogn. Comput. **6**(1), 20 (2022)
115. K. Wang et al., Context sensing attention network for video-based person re-identification. ACM Trans. Multimed. Comput. Commun. Appl. **19**(4), 1–20 (2023)
116. H. Tao, Q. Duan, J. An. An adaptive interference removal framework for video person re-identification. IEEE Trans. Circuits Syst. Video Technol. **33**(9) (2023)
117. S. Bai et al., Salient-to-broad transition for video person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio*n (2022)
118. H.K. Vydana et al., Jointly trained transformers models for spoken language translation, in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2021)
119. X. Liu et al., Very deep transformers for neural machine translation. arXiv preprint http://arxiv.org/abs/2008.07772 (2020)
120. D. Neimark et al., Video transformer network. arXiv preprint http://arxiv.org/abs/2102.00719 (2021)
121. S. He et al., Transreid: transformer-based object re-identification. arXiv preprint http://arxiv.org/abs/2102.04378 (2021)
122. F. Shen et al., GiT: graph interactive transformer for vehicle re-identification. arXiv preprint http://arxiv.org/abs/2107.05475 (2021)
123. S. Khan et al., Transformers in vision: a survey. arXiv preprint http://arxiv.org/abs/2101.01169 (2021)
124. G. Zhang et al., HAT: hierarchical aggregation transformers for person re-identification. arXiv preprint http://arxiv.org/abs/2107.05946 (2021)
125. C. Sharma, S.R. Kapil, D. Chapman, Person re-identification with a locally aware transformer, arXiv preprint http://arxiv.org/abs/2106.03720 (2021)
126. X. Liu et al. A video is worth three views: trigeminal transformers for video-based person re-identification. arXiv preprint http://arxiv.org/abs/2104.01745 (2021)
127. P.K. Sarker, Q. Zhao, M.K. Uddin, Transformer-based person re-identification: a comprehensive review. IEEE Trans. Intell. Veh. (2024). https://doi.org/10.1109/TIV.2024.3350669
128. F. Yang et al., Spatiotemporal interaction transformer network for video-based person re-identification in internet of things. IEEE Internet Things J. **10**(14) (2023)
129. Z. Tang et al., Multi-stage spatio–temporal aggregation transformer for video person re-identification. IEEE Trans. Multimed. **25** (2022). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9996153
130. T. Zhang et al., Spatiotemporal transformer for video-based person re-identification arXiv preprint http://arxiv.org/abs/2103.16469 (2021)
131. X. Zang, G. Li, W. Gao, Multi-direction and multi-scale pyramid in transformer for video-based pedestrian retrieval. IEEE Trans. Ind. Inform. **18**(12) (2022)
132. T. He et al., Dense interaction learning for video-based person re-identification, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021)
133. A. Alsehaim, T.P. Breckon, VID-trans-ReID: enhanced video transformers for person re-identification (2022). https://bmvc2022.mpi-inf.mpg.de/0342.pdf
134. X. Yang et al., STFE: a comprehensive video-based person re-identification network based on spatio–temporal feature enhancement. IEEE Trans. Multimed. (2024). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10420506
135. L. Courtney, R. Sreenivas, Learning from videos with deep convolutional LSTM networks. arXiv preprint http://arxiv.org/abs/1904.04817 (2019)
136. L. Wu, C. Shen, A. van den Hengel, Convolutional LSTM networks for video-based person re-identification. arXiv preprint http://arxiv.org/abs/1606.01609 (2016)
137. S. Li, W. Liu, H. Ma, Attentive spatial–temporal summary networks for feature learning in irregular gait recognition. IEEE Trans. Multimed. **21**(9), 2361–2375 (2019)

138. D. Ouyang, Y. Zhang, J. Shao, Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks. Pattern Recognit. Lett. **117**, 153–160 (2019)
139. D. Avola et al., Bodyprint—a meta-feature based LSTM hashing model for person re-identification. Sensors **20**(18), 5365 (2020)
140. W. Song et al., Extended global local representation learning for video person re-identification. IEEE Access **7**, 122684–122696 (2019)
141. J. Dai et al., Video person re-identification by temporal residual learning. IEEE Trans. Image Process. **28**(3), 1366–1377 (2018)
142. P. Limcharoen, N. Khamsemanan, C. Nattee, Gait recognition and re-identification based on regional LSTM for 2-second walks. IEEE Access **9**, 112057–112068 (2021)
143. A. Bhuiyan, J. Huang, STCA: utilizing a spatio-temporal cross-attention network for enhancing video person re-identification. Image Vis. Comput. **123**, 104474 (2022)
144. W. Xing, Y. Li, S. Zhang, View-invariant gait recognition method by three-dimensional convolutional neural network. J. Electron. Imaging **27**(1), 013010 (2018)
145. G. Zou et al., Person re-identification based on metric learning: a survey. Multimed. Tools Appl. **80**, 26855–26888 (2021)
146. M. Koestinger et al., Large scale metric learning from equivalence constraints, in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012)
147. S. Pedagadi et al., Local fisher discriminant analysis for pedestrian re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013)
148. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)
149. R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2 (IEEE, 2006)
150. F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015)
151. P. Fang et al., Set augmented triplet loss for video person re-identification, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021)
152. J. Wang et al., Deep ranking model by large adaptive margin learning for person re-identification. Pattern Recognit. **74**, 241–252 (2018)
153. J. Yang et al., Spatial–temporal graph convolutional network for video-based person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
154. F. Zhu et al., A loss combination based deep model for person re-identification. Multimed. Tools Appl. **77**(3), 3049–3069 (2018)
155. N. Wojke, A. Bewley, Deep cosine metric learning for person re-identification, in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2018)
156. A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification. arXiv preprint http://arxiv.org/abs/1703.07737 (2017)
157. J. Meng et al., Deep graph metric learning for weakly supervised person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 6074–6093 (2021)
158. L. An, X. Chen, S. Yang, X. Li, Person re-identification by multi-hypergraph fusion. IEEE Trans. Neural Netw. Learn. Syst. **28**(11), 2763–2774 (2017)
159. M. Li, X. Zhu, S. Gong, Unsupervised tracklet person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1770–1782 (2019)
160. M. Li, X. Zhu, S. Gong, Unsupervised person re-identification by deep learning tracklet association, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018)
161. Y. Chen, Zhu, X., Gong, S. Deep association learning for unsupervised video person re-identification. arXiv preprint http://arxiv.org/abs/1808.07301 (2018)
162. M. Ye et al., Dynamic graph co-matching for unsupervised video-based person re-identification. IEEE Trans. Image Process. **28**(6), 2976–2990 (2019)
163. M.V.N.K. Prasad, R. Balakrishnan, Spatio–temporal association rule based deep annotation-free clustering (STARDAC) for unsupervised person re-identification. Pattern Recognit. **122**, 108287 (2022)
164. X. Li et al., Multi-granularity pseudo-label collaboration for unsupervised person re-identification. Comput. Vis. Image Underst. **227**, 103616 (2023)
165. G. Zhang et al., Camera contrast learning for unsupervised person re-identification. IEEE Trans. Circuits Syst. Video Technol. **33**(8) (2023)
166. M. Kim, M. Cho, S. Lee, Feature disentanglement learning with switching and aggregation for video-based person re-identification, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023)
167. Y. Yang et al., Progressive unsupervised video person re-identification with accumulative motion and tracklet spatial–temporal correlation. Future Gener. Comput. Syst. **142**, 90–100 (2023)
168. S. Zeng et al., Anchor association learning for unsupervised video person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **35**(1) (2022)
169. X. Lin et al., Unsupervised person re-identification: a systematic survey of challenges and solutions. arXiv preprint http://arxiv.org/abs/2109.06057 (2021)
170. P. Xie et al., Sampling and re-weighting: towards diverse frame aware unsupervised video person re-identification. IEEE Trans. Multimed. **24**, 4250–4261 (2022)
171. Y. Lin et al., Unsupervised person re-identification via softened similarity learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
172. X. Wang et al., Exploiting global camera network constraints for unsupervised video person re-identification. IEEE Trans. Circuits Syst. Video Technol. **31**(10), 4020–4030 (2020)

173. Xie, P., et al. Unsupervised video person re-identification via noise and hard frame aware clustering, in *2021 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2021)
174. Y. Yan et al., Exploring visual context for weakly supervised person search. arXiv preprint http://arxiv.org/abs/2106.10506 (2021)
175. K.K. Singh et al., Hide-and-seek: a data augmentation technique for weakly-supervised localization and beyond. arXiv preprint http://arxiv.org/abs/1811.02545 (2018)
176. X. Wang et al., Learning person re-identification models from videos with weak supervision. iEEE Trans. Image Process. **30**, 3017–3028 (2021)
177. J. Meng, S. Wu, W.-S. Zheng, Weakly supervised person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019)
178. H.-X. Yu, W.-S. Zheng, Weakly supervised discriminative feature learning with state information for person identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
179. M. Liu et al., Weakly supervised tracklet association learning with video labels for person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
180. M. Liu et al., Iterative local-global collaboration learning towards one-shot video person re-identification. IEEE Trans. Image Process. **29**, 9360–9372 (2020)
181. J. Shao, X. Ma, Hierarchical pseudo-label learning for one-shot person re-identification. Appl. Intell. **52**, 9225–9238 (2022)
182. J. Zhang, N. Wang, L. Zhang, Multi-shot pedestrian re-identification via sequential decision making, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
183. M. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models. Adv. Neural Inf. Process. Systems **23**, 1189–1197 (2010) https://proceedings.neurips.cc/paper_files/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Metadata.json
184. W. Zhang et al., Feature aggregation with reinforcement learning for video-based person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **30**(12), 3847–3852 (2019)
185. M. Krichen, Generative adversarial networks, in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (IEEE, 2023)
186. N.K.S. Behera et al., Person re-identification: a taxonomic survey and the path ahead. Image Vis. Comput. **122**, 104432 (2022)
187. F. Wan et al., When person re-identification meets changing clothes, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2020)
188. X. Qian et al., Long-term cloth-changing person re-identification, in *Proceedings of the Asian Conference on Computer Vision* (2020)
189. C. Dai, C. Peng, M. Chen, Selective transfer cycle GAN for unsupervised person re-identification. Multimed. Tools Appl. **79**, 12597–12613 (2020)
190. Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, in *Proceedings of the IEEE International Conference on Computer Vision* (2017)
191. Z. Zhao et al., JoT-GAN: a framework for jointly training GAN and person re-identification model. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **18**(1s), 1–18 (2022)
192. Y. Chen et al., ResT-ReID: transformer block-based residual learning for person re-identification. Pattern Recognit. Lett. **157**, 90–96 (2022)
193. H. Wang et al., NFormer: robust person re-identification with neighbor transformer. arXiv preprint http://arxiv.org/abs/2204.09331 (2022)
194. M. Zhou et al., Motion-aware transformer for occluded person re-identification. arXiv preprint http://arxiv.org/abs/2202.04243 (2022)

Publisher's Note