Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Deep Learning for Video-based Person Re-Identification: A Survey

Khawar Islam[a,**]

[a]floppydisk.ai, Karachi, Pakistan

ABSTRACT

Video-based person re-identification (video re-ID) has lately fascinated growing attention due to its broad practical applications in various areas, such as surveillance, smart city, and public safety. Nevertheless, video re-ID is quite difficult and is an ongoing stage due to numerous uncertain challenges such as viewpoint, occlusion, pose variation, and uncertain video sequence, etc. In the last couple of years, deep learning on video re-ID has continuously achieved surprising results on public datasets, with various approaches being developed to handle diverse problems in video re-ID. Compared to image-based re-ID, video re-ID is much more challenging and complex. To encourage future research and challenges, this first comprehensive paper introduces a review of up-to-date advancements in deep learning approaches for video re-ID. It broadly covers three important aspects, including brief video re-ID methods with their limitations, major milestones with technical challenges, and architectural design. It offers comparative performance analysis on various available datasets, guidance to improve video re-ID with valuable thoughts, and exciting research directions.

1. Introduction

With the endless efforts of computer vision and deep learning researchers, deep learning has accomplished exceptional success in person re-ID. In a few years, deep learning shows remarkable results in video re-ID and gives new birth to surveillance systems. With the rapid improvement in multimedia technology, video re-ID has gained much more attention in academia and the industrial sector over the last ten years Zheng et al. (2016b); Nambiar et al. (2019). The dominant reason for video re-ID popularity is to provide a wide range of services for public safety such as tracking each person with a unique ID, preventing crimes, behavior analysis, forensic investigation, etc. Almasawa et al. (2019). In intelligent video surveillance applications, video re-ID is defined as recognizing an individual person through various non-overlapping cameras from the huge number of gallery images Chen et al. (2020a). It is one of the intriguing computer vision problems that are present among inter-camera variance challenges such as background clutter, occlusion, viewpoint, illumination changes, human pose variation and etc.

Video re-ID is an extended way of image-based person re-ID. Rather than comparing image pairs, pairs of video sequences are provided to the re-ID algorithm. The essential and important task of the video re-ID algorithm is to obtain temporal features from video sequences. Compare with image-based information, videos naturally comprise more information and evidence than individual images. Lately, numerous methods have been developed for video re-ID Zhou et al. (2017); Zhang et al. (2017). Most existing approaches emphasize extracting spatial and temporal features present in a video and then applying the re-ID algorithm to obtained features. In general, taking a video from different surveillance cameras like CCTV from different outside places. Then, detect persons in a video sequence and create a bounding box on it. Due to the high volume of data, it is difficult to draw manually bounding boxes and annotate each person's image for training.

Different studies Ren et al. (2015); Li et al. (2017); Lan et al. (2018) trained detectors to detect persons in a video sequence. Next, training a new re-ID model on highly noisy data based on previously annotated data. At last, giving query (probe) person image to re-ID model to find query person in a large set of candidate gallery Ye et al. (2021). The main role of video re-ID is to extract spatiotemporal features from video sequences. Some previous studies directly utilized person re-ID methods

**Corresponding author: Khawar Islam
e-mail: khawar512@gmail.com (Khawar Islam)

Table 1. Comparison between existing survey papers and our survey paper. Our survey paper mainly focuses on video re-ID.

Survey	Focus	Major Contribution	Video re-ID	Publication
Mazzon et al. (2012)	Crowd	Drawbacks of existing approaches Proposed simple knowledge-based method	Partial	PRL
Satta (2013)	Appearance Descriptors	Covers public datasets with current evaluation Raised open and closed set re-ID scenarios	Partial	arXiv
Gala and Shah (2014)	Open-Closed	Highlighted public datasets with current evaluation Raised open and closed set re-ID scenarios	Partial	IVC
Zheng et al. (2016b)	Image & Video	Discussed history and relationship of person re-ID Hand-crafted and DL methods are reviewed	Partial	arXiv
Lavi et al. (2018)	re-ID	Survey on deep neural networks techniques Covers loss function and data augmentation	X	arXiv
Wang et al. (2018a)	re-ID	Traditional methods and architectural perspectives CNN, RNN and GAN for person Re-ID	X	CAAI-TIT
Wu et al. (2019)	Image	Survey of SOTA methods with feature designing Several results on ResNet and Inception	P	Neuro-computing
Masson et al. (2019)	Image	Extensively covered pruning methods, strategies Performance evaluation on different datasets	X	JIVP
Nambiar et al. (2019)	Gait	Covered bio-metric details, pose analysis Datasets and Multi-dimensional gait	X	ACM-CS
Leng et al. (2019)	Open-world	Generalized open-world re-ID Specific application driven re-ID	P	IEEE-TCSVT
Wang et al. (2019b)	Heterogeneous	Focused on heterogeneous re-ID Problem of inter-modality discrepancies	Partial	IJCAI
Wang et al. (2020)	Image & Video	Extensive review of previous Re-ID methods Briefly discussed CNN, RNN and GAN	X	IEEE-Access
Ye et al. (2021)	Image&Video	Discussed closed-world and open-world re-ID Baseline for single-/cross-modality re-ID	Partial	IEEE-PAMI
Xiangtan et al. (2021)	Text & Image	Extensively reviewed person search methods Feature learning and identity-driven methods	X	IJCAI
Lin et al. (2021)	Image & Video	Extensively covered unsupervised methods Discussion about dataset and evaluation Peformance analysis and metrics	Partial	arXiV
Ours	**Video**	**Briefly discuss video re-ID methods** **Discuss unique architectures, loss functions** **Performance analysis of current methods**	**Full**	**CVIU**

for images with some extension and applied for video. These approaches extract spatiotemporal information from each image independently by utilizing a recurrent neural network, feature aggregation function, and different pooling operations to obtain a frame-level information (e.g. appearance) representation. These above-mentioned techniques view different video frames with equal importance when needed frame-level features. However, these approaches extract abstract-level global features from the human body, while ignoring several local visual cues from a body such as a gait, hairs and etc.

Person re-ID in videos taken by multiple non-overlapping cameras which is a more practical implementation than images and achieves growing research trends Wang et al. (2014); Zhou et al. (2017). In practical terms, videos captured from surveillance cameras with the involvement of pedestrians are the actual videos for person re-ID because these videos contain useful abundant information and spatial temporal features of a pedestrian that includes different human poses with diverse view angles. Nevertheless, recognizing discriminative portions of pedestrians against noisy data and extracting their features is an intriguing vision problem that is complicated for matching persons. Several video re-ID methods McLaughlin et al. (2017); Zhang et al. (2017) utilize CNN and RNN networks to extract spatio-temporal features from images and employ a pooling strategy to aggregate them. However, following these procedures, the task of matching persons becomes more sensitive when there are some noisy samples in data due to cluttered background or occlusion. While comparing two images of a person, each frame contributes equally to the matching task. For instance, if two persons are occupied with the same occluded object, the same appearance on occluded objects gives a false positive result in person re-ID.

1.1. Contribution of this survey paper

Most of the researchers focused on and surveyed traditional re-ID methods. Several survey papers covered conventional

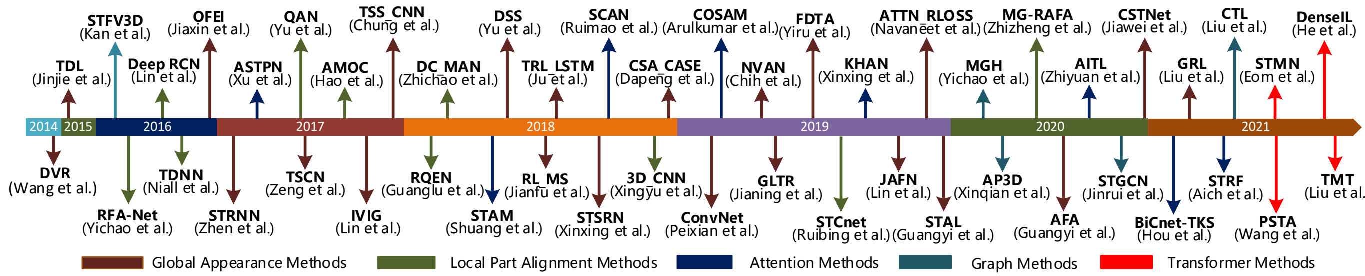

Fig. 1. Timeline of the top-performing methods for video re-ID task.

techniques including feature learning and distance learning, and some of them broadly covered deep learning techniques for re-ID. As far as our deep analysis, there is no survey paper discussing the recent video re-ID methods, novel loss functions, architectural designs, and approaches for video re-ID perspective. In this paper, we discuss comprehensive recent methods published in top-tier conferences and journals. In a nutshell, the contributions discussed in this survey paper are summarized as follows:

1. To the best of our knowledge, this is the very first review paper to extensively cover deep learning methods for video re-ID instead of all types of person re-ID compared with recent existing surveys Ye et al. (2021); Wu et al. (2019); Almasawa et al. (2019).
2. We comprehensively cover deep learning techniques for video re-ID from multiple aspects, including global appearance methods, local part alignment methods, attention methods, graph methods, and transformer methods.
3. This survey paper broadly covers architectural designs, novel loss functions, existing work, and the rapid progress of deep learning for video re-ID. Thus, it gives the readers to overlook the entire video re-ID work.
4. Extensive comparison of top-ranking results on the benchmark datasets is performed. The development of video Re-ID and the challenges affecting video Re-ID systems are discussed, and a brief review and future discussion are given.

1.2. Review of existing survey papers

Our recent survey offers a comprehensive and in-depth review of video re-ID and is distinct from previous and existing surveys and studies, as we broadly include the particular area of intelligent video surveillance and its practical usage. A detailed literature of previous survey articles aiming at person re-ID and its deep learning-based approaches is present, and some of them focused on open-world and close-world re-ID. Still, as far as we know, there is no previous paper that deeply focuses on video-based person re-ID from a practical point of view. We classify the previous and existing work of literature on person re-ID into five major categories named re-ID in the crowd, DL for re-ID, appearance descriptor for re-ID, opened the world, and closed re-ID. The comprehensive application scenario, relevant contributions, and special considered factors about past survey papers are described in Table 1. The influential work on person re-ID applications is mentioned in Zheng et al. (2016b); Almasawa et al. (2019).

Mazzon et al. (2012) presented a state-of-the-art (SOTA) re-ID framework for crowd applications and implementation of the practical framework in crowded scenes where people's movement captured from body appearance and comprehensively covered the discussion about person re-ID applications in terms of appearance features (color, texture, shape), association (distance, learning, optim) and calibration (color, spatial-temp). Similarly, Riccardo Satta (2013), provided a comprehensive overview of appearance descriptors and challenging issues i.e., illumination changes, partial occlusion, changes in color response and pose and viewpoint variations. Furthermore, they covered global and local features with some "other cues". Furthermore, the author in Gala and Shah (2014) broadly discussed person re-ID problems, especially challenges related to the system level and component level. The authors discussed possible re-ID scenarios with comprehensively covered public datasets, evaluation metrics, and current work in re-ID.

One of the major and remarkable surveys Zheng et al. (2016b) focused on each aspect of re-ID and connected with instance retrieval and image classification. A brief milestone and technical perspectives of image/video-based re-ID with handcrafted methods were discussed. Traditional methods for person re-ID Wang et al. (2018a) were highlighted with further extended deep learning approaches such as CNN, RNN, and GAN to achieve person re-ID task and covered advantages and disadvantages. Similarly, the author in Lavi et al. (2018) briefly discussed person re-ID from a video surveillance perspective and covered specific novel re-ID loss functions. The authors provided detailed re-ID approaches and divided them into i.e., recognized, verified deep model, distance learning metrics, feature learning, video-based person re-ID models, and data augmentation models. Further, they also conducted some experiments on the base model with several people re-ID methods Wu et al. (2019). In Masson et al. (2019), a detailed analysis of pruning techniques for compressing re-ID models is presented. To strengthen person the work, the authors further experimented with different pruning techniques and applied them to the deep siamese neural network. Their finding shows that pruning methods substantially decrease the number of parameters without decreasing accuracy.

Different from previous surveys, a gait-based Nambiar et al. (2019) person re-ID has been discussed and highlighted various biometric regions in the human body i.e., hard biometrics including face identity, fingerprint, DNA, eye retina, and palmprint. Similarly, soft biometrics are related to body measurement, eye color, gait, and hair/beard/mustache. Particularly, the authors in Almasawa et al. (2019) briefly discussed traditional and deep learning-based popular architectures and categories into image and video re-ID. Additionally, they compared to rank 1 results with SOTA methods and highlighted important challenges with future direction. Mostly person re-ID systems designed for closed-world settings, in Leng et al. (2019), the authors focused on open-world settings and discussed new trends of person re-ID in that area. They analyzed inconsistencies between open and closed-world applications and briefly discussed data-driven methods. Several specific surveys Mazzon et al. (2012); Nambiar et al. (2019); Wang et al. (2019b) presented a depth literature review of some particular areas like heterogeneous re-ID Wang et al. (2019b), the author studied the concept of Hetero re-ID. They provided a comprehensive literature review in infrared images, low-resolution images, text, and sketches. Afterward, the authors analyzed various datasets with evaluation metrics by giving future insights and providing new challenges areas in Hetero re-ID.

Recently, the author Ye et al. (2021) conducted an extensive literature review of deep learning-based re-ID. Instead of focusing on an overview, they briefly covered limitations and advantages. The new AGW baseline is designed with a novel evaluation metric (mINP) for single and cross-modality re-ID tasks. However, the above survey papers of all presented covered person re-ID surveys do not focus on recent methods of VID re-ID and their solutions for intelligent video surveillance and practical applications. Precisely, we cover recent novel loss functions designed for video re-ID, architectural design, brief technical aspects of significant papers, and broadly discuss performance analysis with the most frequent datasets used for video-based re-ID. Several popular methods are illustrated in Fig. 1

2. Video re-ID Methods

This section discusses the feature representation learning approaches for video re-ID. We divide it into five main categories: a) Global Appearance Methods (subsection 2.1) b) Local Part Alignments Methods (subsection 2.2) c) Attention Methods (subsection 2.3) d) Graphs Methods (subsection 2.4) and e) Transformers Methods (subsection 2.5).

2.1. Global Appearance Methods

This class of methods extracts a single feature vector from a person's image without any supplementary information. Since person re-ID is originally applied for person retrieval problems Zhang et al. (2020a), learning global feature is often ignored in previous studies when incorporating existing DL approaches into the video re-ID domain. As a pioneering work, Niall et al. (2016) introduces the first **Recurrent Deep Neural Network (RDNN)** architecture based on pooling and re-currency mechanism to combine all time-step data into a single feature vector.

To compare different temporal modeling methods, Gao and Nevatia (2018) comprehensively study 3D ConvNET, RNN, temporal pooling, and temporal attention by fixed baseline architecture trained with triplet and softmax cross-entropy losses. Fu et al. (2019) address large-scale video re-ID problem by introducing their **Spatial Temporal Attention (STA)** method. Rather than extracting direct frame-level clues by using average pooling, a 2D ST map is used to measure clip-level feature representation without any additional clues. Generally, features extracted from a single frame contain a lot of noise, illumination, occlusion, and different postures. This results in the loss of discriminative information (e.g., appearance and motion). **Refining Recurrent Unit (RRU)** Liu et al. (2019b) recovers the missing parts with the help of motion context and appearance from the previous frame.

Another popular solution is to explicitly handle alignment problem corruption using occluded regions. Li et al. (2018) employs a unique diversity regularization expression formulated on Hellinger distance to verify the SA models which do not find similar body parts. Zhao et al. (2019) propose an attribute-based technique for feature re-weighting frame and disentanglement. Single frame features are divided into different categories of sub-features, and each category defines a specific semantic attribute. A two-stream network Song et al. (2019) that jointly handle detailed and holistic features utilize an attention approach to extract feature at the global level. Another network captures local features from the video and enhances the discriminative ST features by combining these two features.

Different from Zhang et al. (2020b), a **Global-guided Reciprocal Learning (GRL)** framework Liu et al. (2021d) extracts fine-grained information in an image sequence. Based on local and global features, **Global-guided Correlation Estimation (GCE)** module generates feature correlation maps, locating low and high correlation regions to identify similar persons. Further, to handle multiple memory units and enhance temporal features, **Temporal Reciprocal Learning (TRL)** is constructed to gather specific clues. Li et al. (2021) improve the global appearance by jointly investigating global and local region alignments by considering inter-frame relations.

2.2. Local Part Alignments Methods

These methods extract local part/region that effectively prevents misalignment with other frames in a tracklet. Considering the persistent body structure with the combination of inconsistent body parts in terms of appearance, they are new to each other. The goal is to distinguish personal images based on visual similarity.

To preserve structural relationship details, the **Structural Relationship Learning (SRL)** Bao et al. (2019) is proposed to extract structural relations in a refined and efficient way. SRL helps convolutional features to make the relation useful between regions and GCN. GCN allows learning the feature representations of the hidden layers which encode node features and local structural information of graph. Another popular solution is **Spatial-Temporal Completion network (STCnet)** Hou et al. (2019), a method explicitly handles partial occlusion by recovering the occluded part appearance. **Region-based Quality Estimation Network (RQEN)** Song et al. (2018b) designs an end-to-end training technique with gradient and learns the partial quality of each person image and aggregates complementary partial details of video frames in a sequence.

Different from previous methods, they utilize erasing techniques to penalize regularized terms during network training to prevent over-fitting. Hou et al. (2020) capture complementary affinities from video frames using an erasing strategy during training and testing. Based on the activated parts of previous frames, this approach erases the regions of each frame which ensures the frame concentrate on a new human part. To extract fine-grained cues, **Multi-Granularity Reference aided Attentive Feature Aggregation (MG-RAFA)** is proposed in Zhang et al. (2020b) to jointly handle Spatio-temporal features. Semantic hierarchy is considered for each node/position from a global point of view. For the position of each feature, local affinities are utilized with reference to feature nodes which provide the global structural and appearance information to support different weights to local features. Li et al. (2021) considers a holistic feature for visual similarity of video frames while focusing on the quality that allows the recovery of misaligned parts.

2.3. Attention Methods

These methods usually ignore dissimilar pixels in training and prediction, employing similar pixels to make computational-friendly networks.

Song et al. (2018a) introduce a mask-guided network, where binary masks are used to coexist with corresponding person images to decrease background clutter. Similar to the prior work, Subramaniam et al. (2019), CO-Segmentation approaches have shown remarkable improvements in video re-ID over different baselines by integrating a **Cosegmentation-based Attention (COSAM)** Subramaniam et al. (2021) block among different layers in CNN networks. These CO-segmentation methods are able to extract unique features between person images and use them for channel and spatial-wise attention. In a different work in video re-ID, Chen et al. (2019a) learn spatial-temporal features and calculate an attention score map to specify the quality of different components of a person.

In real-world applications, the motion patterns of humans are the dominant part of re-ID. The Flow Guided-Attention network Kiran et al. (2021) is designed to fuse images and sequence of optical flow using CNN feature extractor which allows encoding of temporal data among spatial appearance information. The Flow Guided-Attention depends on joint SA between optical flow and features to take out unique features among them. Additionally, an approach to aggregate features is proposed for longer input streams for improved representation of video sequences.

Several studies focus on multi-grained and multi-attention approaches to concentrate on important parts of the human body. Hu et al. (2020) introduce **Concentrated Multi-grained Multi-Attention Network (CMMANet)**, multi-attention blocks are proposed to obtain multi-grained details by processing intermediate multi-scale features. Moreover, multiple-attention sub-modules in multi-attention blocks can automatically discover multiple discriminative regions in the frame sequence. Relevant to multi-branch networks, Hou et al. (2021) propose an innovative and computational-friendly video re-ID network that differs from the existing frameworks. **Bilateral Complementary Network (BiCnet)** preserves spatial features from the original image and down-sampling approach to broaden receptive fields and **Temporal Kernel Selection (TKS)** module captures the temporal relationship of videos. Different from previous studies, Chen et al. (2020a) introduces an end-to-end 3D framework to capture salient features of pedestrians in spatial-temporal domains. In this framework, salient 3D bins are selected with the help of two-stream networks and an RNN model to extract motion and appearance information.

2.4. Graph Methods

After the remarkable success of the CNN model Krizhevsky et al. (2012) in image understanding and reconstruction, academic and industrial researchers have focused on developing convolutional approaches for graph data. Recently, researchers combine re-ID methods with graph models and explore Video re-ID Yan et al. (2016). Cheng et al. (2018) develop a training network that jointly handles conventional triplet and contrastive losses through a joint laplacian form that can take complete benefit of ranking data and relationships between training samples. In Shen et al. (2018), a novel unsupervised algorithm is formulated, which maps the ranking mechanism in the person re-ID method. Then, the formulation procedure is extended to be able to utilize ranking results from multiple algorithms. Only matching scores produced by different algorithms can lead to consensus results. The key role of the person re-ID task is to efficiently calculate the visual resemblance among person images. Still, ongoing person re-ID methods usually calculate the similarity of different image pairs (investigated) and candidate lists separately whereas neglecting the association knowledge among various query-candidate pairs.

To solve the above problems, **Similarity-Guided Graph Neural Network (SGGNN)** Chen et al. (2018b) propose to generate a graph to illustrate the pairwise associations between query and candidate pairs (nodes) and utilize these associations to provide up-to-date query candidate correlation features extracted from the image in an end-to-end manner. Most re-ID approaches emphasize local features for similarity matching. Chen et al. (2018b) combine multiple person images to estimate the association between local relation and global relation in their **Conditional Random Field (CRF)**. The benefit of this model is to learn local resemblance metrics from image pairs whereas considering the dependencies of all images in a collection, shaping group similarities. Yan et al. (2019) put more effort into person re-ID and employ context details. They first develop a contextual module called the instance expansion part, which emphasizes on relative attention part to find and purify beneficial context detail in the scene. One of the innovative works Wu et al. (2020) for video re-ID is graph-based adaptive representation. Existing studies ignore part-based features, which contain temporal and spatial information. This approach allows the association between contextual information and relevant regional features such as feature affinity and poses alignment connection, to propose an adaptive structure-aware contagiousness graph. Liu et al. (2021b) present **Correlation and**

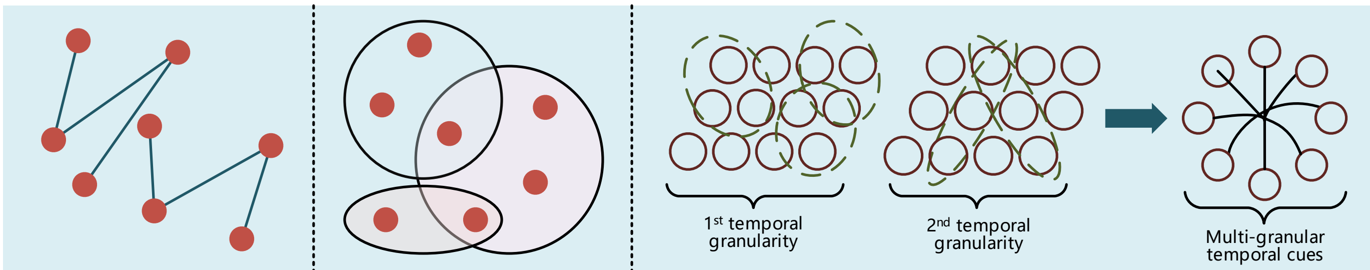

Fig. 2. Illustration of simple graph each line connected with two vertices. In a hypergraph, each edge is connected with more than two vertices. In a multi-granularity graph, each node models specific spatial granularity, and each hypergraph is connected with multiple nodes.

Topology Learning (CTL) method which generates robust and distinguish features. It captures features at multi-granularity levels and overcomes posing appearance problems.

Lately, hyper GNNs have attracted a lot of attention and achieved dominant results in various computer vision research fields such as person re-ID Shen et al. (2018), action recognition Wang and Gupta (2018) and image recognition Chen et al. (2019b). These hypergraph algorithms develop pairwise relationships on the basis of object interest. In general, a hypergraph is a graph in which edges independently work and can join any considerable number of vertices. The illustration of hypergraph as shown in Fig. 2 (b) Conversely, as represented in Fig. 2 (a) where an edge exactly links with two vertices in a simple graph. In MG hypergraph, as represented in Fig. 2 (d) hypergraphs with distinct spatio granularities are built utilizing numerous stages of features like body part throughout the video frames. In every hypergraph stage, novel temporal granularities are taken by hyperedges which connect a type of nodes in a graph such as body part features around separate temporal scales. The first **Multi-Granular Hypergraph (MGH)** Yan et al. (2020) hypergraph and innovative mutual information loss function are proposed to overcome the image retrieval problem. The MGH approach clearly supports multi-granular ST information from the frame sequence. Then, they propose an attention procedure to combine features presenting at the node level to obtain better discriminative graph representations. Remarkably, the proposed approach achieves 90% rank-1 accuracy which is amongst the highest accuracy on the MARS dataset. Label estimation in graph matching is closely related to person re-ID problems in unsupervised learning. Ye et al. (2019) present an unsupervised **Dynamic Graph Matching (DGM)** video re-ID approach to predict labels. This technique iterates the update process by utilizing a discriminative metric and correspondingly updated labels.

2.5. Transformer Methods

Recently, transformer shows a great interest in the computer vision field, and self-attention-based methods are proposed to solve visual problems. Inspired by recent development, Zhang et al. (2021b) put forward the first step, propose first **SpatioTemporal transformer (STT)** and synthesize pre-training data strategy to reduce over-fitting for video re-ID task. In their network, the global module enables supplement to utilize the relation among patches from frames. To extract comprehensive features from videos, Liu et al. (2021c) further explore transformers and introduce **Trigeminal Transformers (TMT)** with robust novel feature extractor that jointly transform raw videos into S, T, and ST domains. To capture fine-grained features and aggregate in multi-view features, a self-view transformer is proposed to enhance single-view features and a cross-view transformer is used to combine multiple features. A **Duplex SpatioTemporal Filtering Network (DSFN)** Zheng et al. (2021) architecture is designed to extract static and dynamic data from frame sequences for video re-ID. To enhance the capability of kernels, sparse-orthogonal constraints are developed to broaden the difference in temporal features. To collaborate with a group of kernels, they add additional channels to assist and extract ST clues from distinct features. A Hybrid **Dense Interaction Learning (DenseIL)** framework is presented in He et al. (2021) which utilizes both CNN and Attention mechanism for video re-ID. DenseIL consists of a CNN-based encoder which is responsible to extract efficient discriminative spatial features and a DI-based decoder densely modeling the ST inherent interaction among frames.

3. Novel Architectures

Different from existing architectures, Jiang et al. (2021) propose a novel design to handle misalignment problems in video re-ID. **Self-Separated network (SSN)** provides an effective approach to deal with temporal and spatial variation of a person's body parts. SSN derives a two-round classification approach, leading to better training in pixel-wise and aggregated features. The improved **Coarse-to-Fine Axial Attention Network (CF-AAN)** Liu et al. (2021a) is designed with the help of Link and re-Detect block which can align noisy tracklist on the image level. This module not only decreases computational costs but also achieves promising results. Various video re-ID methods are still suffering from pose changes and personal misalignment problems. To handle misalignment, Zhang et al. (2021a) propose the **Reference-Aided Part-Aligned (RAPA)** that focuses on different parts of the body and disentangles the discriminative features. **Reference Feature Learning (RFL)** pose-based module is provided to capture uniform standards for alignment. Aligning the body parts in intra-video, relations, and attention-based **Part Feature Disentangling (PFD)** blocks are designed to locate and match body parts through frames.

Most video re-ID methods focus on the important region of the image, therefore, these methods can easily lose out on fine-

Table 2. Training configuration of novel architectures. LR denotes learning rate and L represents loss

Reference and Venue	Method	Extractor	L. Function	LR	Optimizer	Epochs
Wang et al. (2014) ECCV	DVR	HOG3D	Hinge	—	—	—
Karanam et al. (2015) CVPR	SRID	Schmid, Gabor filters	—	—	—	—
Liu et al. (2015) ICCV	STFV3D	Fisher Vector	—	—	—	—
Wu et al. (2016) ARXIV	Deep RCN	—	—	—	—	—
You et al. (2016) CVPR	TDL	HOG3D, Color Histograms, LBP	Hinge	—	—	—
Chen et al. (2016) IEEE-SRL	OFEI	LBP	—	—	—	—
Chen et al. (2016) ECCV	RFA-Net	LBP, HSV, Lab	Softmax	0.001 to 0.0001	—	400
Niall et al. (2016) CVPR	CNN and RNN	Cross Entropy	—	0.001	SGD	500
Zhou et al. (2017) CVPR	JS-TRNN	TAM and SRM	Triplet	—	—	—
Liu et al. (2017) CVPR	QAN	—	Softmax and Triplet	—	—	—
Xu et al. (2017) ICCV	ASTPN	—	CE and Hinge	0.001	SGD	700
Chung et al. (2017) ICCV	2SCNN	CNN and RNN	Softmax	0.001	SGD	1000
Gao et al. (2021) ACM_MM	CMA	CNN+RNN	Softmax	0.001	SGD	800

grained hints in image sequences. Different from previous studies, the novel GRL Liu et al. (2021d) framework is introduced along with reciprocal learning and correlation estimation. The GCE module creates the feature maps of local and global features that helps to locate the low regions and high regions to identify a similar person. Then, a novel TRL approach is introduced to improve the high-correlation semantic information. Gu et al. (2020) propose **Appearance Preserving 3D Convolution (AP3D)** and **Appearance-Preserving Module (APM)**, which align neighborhood feature maps in pixel-level. 3D ConvNets model temporal information on the basis of preserving the quality of visual appearance. It may be easier to aggregate AP3D with current 3DConNet by substituting prior 3D-Conv filters to AP3Ds. In video re-ID, personal attributes and visual appearance are key to matching identities, and both features significantly contribute to the tracking of pedestrians. Novel TAL-Net Liu et al. (2020) is proposed to focus on attribute-temporal learning by constructing a branch network with the help of SA and temporal-semantic context.

4. Loss Functions

Loss function plays a major and crucial role in discriminating the learned features. In general, the softmax loss separates the learned features rather than discriminates. The main goal of designing a person re-ID loss function is to enhance representation with an efficiency loss. We highlight several of the most influential loss functions for video re-ID.

4.1. Attention and CL Loss

Pathak et al. (2020) introduce CL centers online soft mining loss which utilizes center vectors from center loss as class label vector representations to crop out those frames that contain higher noise because it contains high variance compared to the original classifier weights. Additionally, they penalize the model by giving maximum attention scores to those frames that have randomly deleted patches. Those random erased frames are labeled as 1 otherwise 0 and N is the number of total frames.

$$\text{AL} = \frac{1}{N}\sum_{i=1}^{N} \text{label}(i) * \text{ Attention }_{score}(i) \tag{1}$$

4.2. Weighted Triple-Sequence Loss (WTSL)

Jiang et al. (2020) explicitly encode frame-based image level information into video level features that can decrease the effect of outlier frame. Intra-class distance in WTSL makes similar videos closer and inter-class distance pushes dissimilar videos further apart.

$$\text{L}_{WTSL} = \sum_{i=1}^{N}\left[\left\| \text{F}_a^i - \text{F}_p^i\right\|_2^2 - \left\|\text{F}_a^i - \text{F}_n^i\right\|_2^2 + \alpha\right]_2 \tag{2}$$

where α represents margin, N is the number of triple-sequences and P represents person ID. The F_a is a closer feature to its own class centroid and far away from other class centroids.

4.3. Symbolic Triplet Loss (STL)

Aruna Kumar et al. (2020) propose STL which utilizes the Wasserstein metric to overcome the representation problem which allows obtaining the distance between feature vectors that are symbolic.

$$D_w\left(\psi_i, \psi_j\right) = \sum_{m=1}^{M}\sum_{t=1}^{T} \psi_{im}{}^{-1}(t) - \psi_{jm}{}^{-1}(t) \tag{3}$$

where ψ_i and ψ_j denote the distributions of multi-dimensional feature vectors at the ith and jth. $\psi_i{}^{-1}(t)$ is the quantile function and M is the feature of each video.

4.4. Weighted Contrastive Loss (WCL)

Wang et al. (2019a) construct WCL by the combination of traditional contrastive loss. The purpose of this loss function is to allocate an appropriate weight for every proper image pair.

$$L_{WCL}(N) = \frac{1}{2}\frac{\sum_{(x_i,x_j)\in N} w_{ij}^{-} \max\left(0, \alpha - d_{ij}\right)^2}{\sum_{(x_i,x_j)\in N} w_{ij}^{-}} \tag{4}$$

$$L_{WCL}(P,N) = (1-\lambda)L_{WCL}(P) + \lambda L_{WCL}(N) \tag{5}$$

where hyperparameter λ handles the contribution of both positive and negative sets towards final value of contrastive loss.

4.5. Triplet Loss

Chen et al. (2019a) design triplet loss to conserve ranking relationship among videos of pedestrian triplets. In triplet loss, the distance between feature pairs belonging to similar classes decreases, while the distance between feature pairs of different classes increases.

$$\begin{aligned} L_{tri} = & \sum_{i,j,k\in\Omega} \left[d_g(i,j) - d_g(i,k) + m_g\right]_+ \\ & + \sum_{i,j,k\in\Omega} \lambda \left[d_l(i,j) - d_l(i,k) + m_l\right]_+ \end{aligned} \tag{6}$$

where m_g and m_l represent thresholds margin to restrict the distance gap between positive and negative samples and $[x]^+$ is the max function $\max(0,x)$.

4.6. Regressive Pairwise Loss (RPL)

Liu et al. (2018) develop RPL to improve pairwise similarity by combining all positive sets in one single subspace. It helps with the soft margin between positive sets and is harder than the general triplet loss.

$$\begin{aligned} L_p\left(x_i, x_j, y\right) = y \cdot \max\left\{d\left(x_i, x_j\right) - \log(\alpha), 0\right\} \\ +(1-y)\cdot \max\left\{\alpha - d\left(x_i, x_j\right), 0\right\} \end{aligned} \tag{7}$$

where y denotes label whether x_i and x_j are similar people. If a person is from the same identity it is represented as 1 otherwise 0. When y = 0, RPL pushes samples far away from each other beyond margin α. When y = 1, RPL pulls the samples together within distances no more than $\log(\alpha)$.

5. Datasets and Metrics

We first describe the statistics of benchmark datasets that are frequently used for evaluating video re-ID methods. Secondly, we broadly review the performance of previous superior methods in chronological order. Lastly, we analyze results based on several major factors for video re-ID.

5.1. Training and Testing Datasets

Since video re-ID is a real-world problem and closer to a video surveillance scenario. During past years, various demanding datasets have been constructed for video re-ID: MARS Zheng et al. (2016a), DukeMTMC-VID Wu et al. (2018) and iLIDS-VIDWang et al. (2014), these three datasets are commonly used for training and evaluation, because of the large number of track-lets and pedestrian identities.

5.1.1. MARS

The dataset is constructed based on six synchronized CCTV cameras. It comprises 1,261 pedestrians with different varieties of images (poor image quality, poses, colors, and illuminations) captured by two cameras. It is extremely difficult to match pedestrian images because it contains 3,248 distractors to make the dataset more real-world.

5.1.2. DukeMTMC-VID

It is a subgroup of the DukeMTMC dataset which purely consists of 8 cameras with high-resolution images. It is one of the large-scale datasets where pedestrian images are cropped using manual hand-drawn bounding boxes. Overall, it comprises 702 identities, 16,522 training images 17,661 gallery images, and 2,228 probe images.

5.1.3. iLIDS-VID

It is one of the challenging datasets which contains 300 pedestrians captured by two CCTV cameras in public. Due to the public images, it contains lighting, viewpoint changes, different similarities, background clutter, and occlusions. It consists of a 600 sequence of images of 300 diverse individual images. Each sequence of the pedestrian images has a range length of 23 to 192 and the number of frames is 73.

5.2. Evaluation Protocol

The are two standard evaluation protocols for evaluating video re-ID methods which are mAP and CMC. CMC is the probability of top top-K correct matches in a retrieval list. Another evaluation metric is mAP, which measures the average retrieval accuracy with multiple GT.

6. Analysis and Future Direction

We broadly review the top-performing methods from video re-ID perspectives. We mostly focus on the work published in 2018 till now. Specifically, we include STAN Li et al. (2018), Snippet Chen et al. (2018a), STA Fu et al. (2019), ADFD Zhao et al. (2019), VRSTC Hou et al. (2019), GLTR Li et al. (2019), COSAM Subramaniam et al. (2019), STE-NVAN Liu et al. (2019a), MG-RAFA Zhang et al. (2020b), MGH Yan et al. (2020), STGCN Yang et al. (2020), TCLNet Hou et al. (2020), AP3D Gu et al. (2020), AFA Chen et al. (2020b), PSTAWang et al. (2021a) DenseIL He et al. (2021), STMN Eom et al. (2021), STRF Aich et al. (2021), SANet Bai et al. (2021), DPRAM Yang et al. (2021), HMN Wang et al. (2021b), GRL Liu et al. (2021d), and TMT Liu et al. (2021c). We summarize the video re-ID results on three widely used benchmark datasets. Table. 3 highlights the backbone, mAP and R-1 results, and methods.

Firstly, with the recent development of self-attention-based methods, several video re-ID methods have obtained higher mAP and top-1 accuracy (Liu et al. (2021c) 91.2%) on the widely used MARS dataset. Especially, DenseIL He et al. (2021) achieves the highest mAP of 87.0% but rank-1 accuracy is 90.8% which is slightly lower than TMTLiu et al. (2021c) on MARS dataset. The advantage of the DenseIL He et al. (2021) method is to simultaneously use CNN and attention-based architecture to efficiently encode spatial information into discriminative features. Those methods focus on long-range relationships and specific part-level information on an input signal. Various popular methods separately learn weights and spatial-temporal features Hou et al. (2019, 2020). Another observation in Zhang et al. (2021b) illustrates that capturing and aggregating pedestrian cues is spatial-temporal while ignoring discrepancies including background areas, viewpoint, and occlusions.

Table 3. Performance analysis of top-performing approaches on Duke, iLIDS, and Mars datasets. "NL" represents a non-local block.

Method	Backbone	MARS		DukeV		iLIDS
		mAP	R-1	mAP	R-1	R-1
STAN$_{(CVPR'18)}$	Res-50	65.8	82.3	×	×	80.2
Snippet$_{(CVPR'18)}$	Res-50	76.1	86.3	×	×	85.4
STA$_{(AAAI'19)}$	Res-50	80.8	86.3	94.9	96.2	×
ADFD$_{(CVPR'19)}$	Res-50	78.2	87.0	×	×	86.3
VRSTC$_{(CVPR'19)}$	Res-50	82.3	88.5	93.5	95.0	83.4
GLTR$_{(ICCV'19)}$	Res-50	78.5	87.0	93.7	96.3	86.0
COSAM$_{(ICCV'19)}$	SERes-50	79.9	84.9	94.1	95.4	79.6
STE-NVAN$_{(BMVC'19)}$	Res-50-NL	81.2	88.9	93.5	95.2	×
MG-RAFA$_{(CVPR'20)}$	Res-50	85.9	88.8	×	×	88.6
MGH$_{(CVPR'20)}$	Res-50-NL	85.8	90.0	×	×	85.6
STGCN$_{(CVPR'20)}$	Res-50	83.7	90.0	95.7	97.3	×
TCLNet$_{(ECCV'20)}$	Res-50-TCL	85.1	89.8	96.2	96.9	86.6
AP3D$_{(ECCV'20)}$	AP3D	85.1	90.1	95.6	96.3	86.7
AFA$_{(ECCV'20)}$	Res-50	82.9	90.2	95.4	97.2	88.5
HMN$_{(TCSVT'21)}$	Res-50	88.8	89	95.1	96.2	×
SANet$_{(TCSVT'21)}$	Res-50	86.0	91.2	96.7	97.7	×
DPRAM$_{(TIP'21)}$	Res-50	83.0	89.0	95.6	97.1	×
PSTA$_{(ICCV'21)}$	Res-50	85.8	91.5	97.4	98.3	91.5
STRF$_{(ICCV'21)}$	Res-50	86.1	90.3	96.4	97.4	89.3
DenseIL$_{(ICCV'21)}$	Res-50	87.0	90.8	97.1	97.6	92.0
STMN$_{(ICCV'21)}$	Res-50	83.7	89.9	94.6	96.7	80.6
GRL$_{(ICCV'21)}$	Res-50	84.8	91.0	×	×	90.4
TMT$_{(arXiv'21)}$	Res-50	85.8	91.2	×	×	91.3

However, in a real-world scenario, the visual data contains a lot of diverse modalities such as recording information, camera ID, etc. Most studies focus on visual similarity by matching probe images into gallery images. Thus, it neglects textual information which is not a good idea. Proposing a new method that extracts visual-textual information at the same time would be helpful in a real-world environment and it will also help to provide more accurate results.

Secondly, annotating new datasets with accurate labels on different CCTV cameras is an expensive and laborious task. In most cases, annotated data are wrong-labeled due to various factors such as person visibility, background clutter, and noise issues in images. Several researchers focus on unsupervised methods Ye et al. (2018, 2019) and active learning approaches Wang et al. (2018b) to alleviate the annotation problem. Still, the accuracy of unsupervised video re-ID methods degrades significantly compare to supervised video re-ID methods. In the future, introducing a unique video re-ID method that facilitates clustering and label assignment will be considered to improve existing unsupervised methods. Further, designing a specific data augmentation policy in a re-ID search space can easily increase the overall performance for all re-ID methods.

Finally, the accuracy on three challenging datasets reaches a difficult state, where the performance gap is less than 1% accuracy like PSTA Wang et al. (2021a) and DenseIL He et al. (2021) on the DukeVID dataset. As a result, it is still difficult to select the best superior method. On iLIDS, the rank-1 performance of PSTA Wang et al. (2021a) is 91.5% and TMT Liu et al. (2021c) is 91.3%. However, most video re-ID architectures are complex in terms of the number of parameters for learning invariant feature representations on combined datasets. Meanwhile, re-ID methods use metric learning techniques like euclidean distance to calculate feature similarity which is time-consuming and slow retrieval and not applicable in real-world applications. How to design a new strategy to replace metric learning strategies still needs more research. Thus, further exploration of video re-ID approaches remains an interesting area for future research.

7. Conclusion

This paper presents a comprehensive review of global appearance, local part alignment methods, graph learning, attention, and transformer model in video re-ID. We provide specific loss functions with mathematical representation to help new researchers to use them instead of using straightforward common loss functions for video re-ID. Finally, we highlight widely and frequently used datasets for evaluating video re-ID techniques and analyze the performance of different methods and provide future research direction.

References

Aich, A., Zheng, M., Karanam, S., Chen, T., Roy-Chowdhury, A.K., Wu, Z., 2021. Spatio-temporal representation factorization for video-based person re-identification, in: International Conference on Computer Vision (ICCV).

Almasawa, M.O., Elrefaei, L.A., Moria, K., 2019. A survey on deep learning-based person re-identification systems. IEEE Access .

Aruna Kumar, S., Yaghoubi, E., Proença, H., 2020. A symbolic temporal pooling method for video-based person re-identification. arXiv .

Bai, S., Ma, B., Chang, H., Huang, R., Shan, S., Chen, X., 2021. Sanet: Statistic attention network for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology .

Bao, L., Ma, B., Chang, H., Chen, X., 2019. Preserving structural relationships for person re-identification, in: ICMEW, IEEE. pp. 120–125.

Chen, D., Li, H., Xiao, T., Yi, S., Wang, X., 2018a. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Chen, D., Xu, D., Li, H., Sebe, N., Wang, X., 2018b. Group consistent similarity learning via deep crf for person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Chen, G., Lu, J., Yang, M., Zhou, J., 2019a. Spatial-temporal attention-aware learning for video-based person re-identification. IEEE Transactions on Image Processing .

Chen, G., Lu, J., Yang, M., Zhou, J., 2020a. Learning recurrent 3d attention for video-based person re-identification. IEEE IEEE Transactions on Image Processing .

Chen, G., Rao, Y., Lu, J., Zhou, J., 2020b. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification?, in: European Conference on Computer Vision.

Chen, J., Wang, Y., Tang, Y.Y., 2016. Person re-identification by exploiting spatio-temporal cues and multi-view metric learning. IEEE Signal Processing Letters 23, 998–1002.

Chen, Z.M., Wei, X.S., Wang, P., Guo, Y., 2019b. Multi-label image recognition with graph convolutional networks, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Cheng, D., Gong, Y., Chang, X., Shi, W., Hauptmann, A., Zheng, N., 2018. Deep feature learning via structured graph laplacian embedding for person re-identification. PR .

Chung, D., Tahboub, K., Delp, E.J., 2017. A two stream siamese convolutional neural network for person re-identification, in: Proceedings of the IEEE international conference on computer vision, pp. 1983–1991.

Eom, C., Lee, G., Lee, J., Ham, B., 2021. Video-based person re-identification with spatial and temporal memory networks, in: International Conference on Computer Vision (ICCV).

Fu, Y., Wang, X., Wei, Y., Huang, T., 2019. Sta: Spatial-temporal attention for large-scale video-based person re-identification, in: Association for the Advancement of Artificial Intelligence.

Gala, A., Shah, S.K., 2014. A survey of approaches and trends in person re-identification. Image and vision computing 32, 270–286.

Gao, J., Nevatia, R., 2018. Revisiting temporal modeling for video-based person reid. British Machine Vision Conference (BMVC) .

Gao, Z., Shao, Y., Guan, W., Liu, M., Cheng, Z., Chen, S., 2021. A novel patch convolutional neural network for view-based 3d model retrieval, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2699–2707.

Gu, X., Chang, H., Ma, B., Zhang, H., Chen, X., 2020. Appearance-preserving 3d convolution for video-based person re-identification, in: European Conference on Computer Vision.

He, T., Jin, X., Shen, X., Huang, J., Chen, Z., Hua, X.S., 2021. Dense interaction learning for video-based person re-identification. International Conference on Computer Vision (ICCV) .

Hou, R., Chang, H., Ma, B., Huang, R., Shan, S., 2021. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. IEEE / CVF Computer Vision and Pattern Recognition Conference .

Hou, R., Chang, H., Ma, B., Shan, S., Chen, X., 2020. Temporal complementary learning for video person re-identification, in: European Conference on Computer Vision.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X., 2019. Vrstc: Occlusion-free video person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Hu, P., Liu, J., Huang, R., 2020. Concentrated multi-grained multi-attention network for video based person re-identification. arXiv .

Jiang, M., Leng, B., Song, G., Meng, Z., 2020. Weighted triple-sequence loss for video-based person re-identification. Neurocomputing .

Jiang, X., Qiao, Y., Yan, J., Li, Q., Zheng, W., Chen, D., 2021. Ssn3d: Self-separated network to align parts for 3d convolution in video person re-identification, in: Association for the Advancement of Artificial Intelligence.

Karanam, S., Li, Y., Radke, R.J., 2015. Sparse re-id: Block sparsity for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 33–40.

Kiran, M., Bhuiyan, A., Blais-Morin, L.A., Javan, M., Ayed, I.B., Granger, E., 2021. A flow-guided mutual attention network for video-based person re-identification.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems .

Lan, W., Dang, J., Wang, Y., Wang, S., 2018. Pedestrian detection based on yolo network model, in: ICMA.

Lavi, B., Serj, M.F., Ullah, I., 2018. Survey on deep learning techniques for person re-identification task. arXiv .

Leng, Q., Ye, M., Tian, Q., 2019. A survey of open-world person re-identification. IEEE Transactions on Circuits and Systems for Video Technology .

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S., 2017. Scale-aware fast r-cnn for pedestrian detection. TM .

Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S., 2019. Global-local temporal representations for video person re-identification, in: International Conference on Computer Vision (ICCV).

Li, Q., Huang, J., Gong, S., 2021. Local-global associative frame assemble in video re-id. British Machine Vision Conference (BMVC) .

Li, S., Bak, S., Carr, P., Wang, X., 2018. Diversity regularized spatiotemporal attention for video-based person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Lin, X., Ren, P., Yeh, C.H., Yao, L., Song, A., Chang, X., 2021. Unsupervised person re-identification: A systematic survey of challenges and solutions. arXiv preprint arXiv:2109.06057 .

Liu, C.T., Chen, J.C., Chen, C.S., Chien, S.Y., 2021a. Video-based person re-identification without bells and whistles. arXiv .

Liu, C.T., Wu, C.W., Wang, Y.C.F., Chien, S.Y., 2019a. Spatially and temporally efficient non-local attention network for video-based person re-identification. British Machine Vision Conference (BMVC) .

Liu, J., Zha, Z.J., Wu, W., Zheng, K., Sun, Q., 2021b. Spatial-temporal correlation and topology learning for person re-identification in videos. arXiv .

Liu, J., Zhu, X., Zha, Z.J., 2020. Temporal attribute-appearance learning network for video-based person re-identification. arXiv .

Liu, K., Ma, B., Zhang, W., Huang, R., 2015. A spatio-temporal appearance representation for viceo-based pedestrian re-identification, in: Proceedings of the IEEE international conference on computer vision, pp. 3810–3818.

Liu, X., Zhang, P., Yu, C., Lu, H., Qian, X., Yang, X., 2021c. A video is worth three views: Trigeminal transformers for video-based person re-identification. arXiv .

Liu, X., Zhang, P., Yu, C., Lu, H., Yang, X., 2021d. Watching you: Global-guided reciprocal learning for video-based person re-identification, in: International Conference on Computer Vision (ICCV).

Liu, Y., Yan, J., Ouyang, W., 2017. Quality aware network for set to set recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5790–5799.

Liu, Y., Yuan, Z., Zhou, W., Li, H., 2019b. Spatial and temporal mutual promotion for video-based person re-identification, in: Association for the Advancement of Artificial Intelligence.

Liu, Z., Wang, Y., Li, A., 2018. Hierarchical integration of rich features for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology .

Masson, H., Bhuiyan, A., Nguyen-Meidine, L.T., Javan, M., Siva, P., Ayed, I.B., Granger, E., 2019. A survey of pruning methods for efficient person re-identification across domains. arXiv .

Mazzon, R., Tahir, S.F., Cavallaro, A., 2012. Person re-identification in crowd. Pattern Recognition Letters 33, 1828–1837.

McLaughlin, N., del Rincon, J.M., Miller, P., 2017. Video person re-identification for wide area tracking based on recurrent neural networks. IEEE Transactions on Circuits and Systems for Video Technology .

Nambiar, A., Bernardino, A., Nascimento, J.C., 2019. Gait-based person re-identification: A survey. ACM CSUR .

Niall, M., Del Rincon, J.M., Miller, P., 2016. Recurrent convolutional network for video-based person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Pathak, P., Eshratifar, A.E., Gormish, M., 2020. Video person re-id: Fantastic techniques and where to find them (student abstract), in: Association for the

Advancement of Artificial Intelligence.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv .

Satta, R., 2013. Appearance descriptors for person re-identification: a comprehensive review. arXiv preprint arXiv:1307.5748 .

Shen, Y., Li, H., Yi, S., Chen, D., Wang, X., 2018. Person re-identification with deep similarity-guided graph neural network, in: European Conference on Computer Vision.

Song, C., Huang, Y., Ouyang, W., Wang, L., 2018a. Mask-guided contrastive attention model for person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S., 2018b. Region-based quality estimation network for large-scale person re-identification, in: Association for the Advancement of Artificial Intelligence.

Song, W., Wu, Y., Zheng, J., Chen, C., Liu, F., 2019. Extended global–local representation learning for video person re-identification. Access .

Subramaniam, A., Nambiar, A., Mittal, A., 2019. Co-segmentation inspired attention networks for video-based person re-identification, in: International Conference on Computer Vision (ICCV).

Subramaniam, A., Vaidya, J., Ameen, M.A.M., Nambiar, A., Mittal, A., 2021. Co-segmentation inspired attention module for video-based computer vision tasks .

Wang, H., Du, H., Zhao, Y., Yan, J., 2020. A comprehensive overview of person re-identification approaches. Ieee Access 8, 45556–45583.

Wang, K., Wang, H., Liu, M., Xing, X., Han, T., 2018a. Survey on person re-identification based on deep learning. CAAI Transactions on Intelligence Technology 3, 219–227.

Wang, M., Lai, B., Jin, Z., Gong, X., Huang, J., Hua, X., 2018b. Deep active learning for video-based person re-identification. arXiv .

Wang, T., Gong, S., Zhu, X., Wang, S., 2014. Person re-identification by video ranking, in: European Conference on Computer Vision.

Wang, X., Gupta, A., 2018. Videos as space-time region graphs, in: European Conference on Computer Vision.

Wang, X., Hua, Y., Kodirov, E., Hu, G., Robertson, N.M., 2019a. Deep metric learning by online soft mining and class-aware attention, in: Association for the Advancement of Artificial Intelligence.

Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D., 2021a. Pyramid spatial-temporal aggregation for video-based person re-identification, in: International Conference on Computer Vision (ICCV).

Wang, Z., He, L., Tu, X., Zhao, J., Gao, X., Shen, S., Feng, J., 2021b. Robust video-based person re-identification by hierarchical mining. IEEE Transactions on Circuits and Systems for Video Technology .

Wang, Z., Wang, Z., Zheng, Y., Wu, Y., Zeng, W., Satoh, S., 2019b. Beyond intra-modality: A survey of heterogeneous person re-identification. arXiv .

Wu, D., Zheng, S.J., Zhang, X.P., Yuan, C.A., Cheng, F., Zhao, Y., Lin, Y.J., Zhao, Z.Q., Jiang, Y.L., Huang, D.S., 2019. Deep learning-based methods for person re-identification: A comprehensive review. Neurocomputing 337, 354–371.

Wu, L., Shen, C., Hengel, A.v.d., 2016. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. arXiv preprint arXiv:1606.01609 .

Wu, Y., Bourahla, O.E.F., Li, X., Wu, F., Tian, Q., Zhou, X., 2020. Adaptive graph representation learning for video person re-identification. IEEE Transactions on Image Processing 29, 8821–8830.

Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y., 2018. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Xiangtan, L., Ren, P., Xiao, Y., Chang, X., Hauptmann, A., 2021. Person search challenges and solutions: A survey. arXiv preprint arXiv:2105.01605 .

Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P., 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification, in: Proceedings of the IEEE international conference on computer vision, pp. 4733–4742.

Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X., 2016. Person re-identification via recurrent feature aggregation, in: European Conference on Computer Vision.

Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y., Shao, L., 2020. Learning multi-granular hypergraphs for video-based person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Yan, Y., Zhang, Q., Ni, B., Zhang, W., Xu, M., Yang, X., 2019. Learning context graph for person search, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Yang, J., Zheng, W.S., Yang, Q., Chen, Y.C., Tian, Q., 2020. Spatial-temporal graph convolutional network for video-based person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Yang, X., Liu, L., Wang, N., Gao, X., 2021. A two-stream dynamic pyramid representation model for video-based person re-identification. IEEE Transactions on Image Processing .

Ye, M., Lan, X., Yuen, P.C., 2018. Robust anchor embedding for unsupervised video person re-identification in the wild, in: European Conference on Computer Vision, pp. 170–186.

Ye, M., Li, J., Ma, A.J., Zheng, L., Yuen, P.C., 2019. Dynamic graph co-matching for unsupervised video-based person re-identification. IEEE Transactions on Image Processing 28, 2976–2990.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C., 2021. Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence .

You, J., Wu, A., Li, X., Zheng, W.S., 2016. Top-push video-based person re-identification, in: Proceedings of the ieee conference on computer vision and pattern recognition, pp. 1345–1353.

Zhang, G., Chen, Y., Dai, Y., Zheng, Y., Wu, Y., 2021a. Reference-aided part-aligned feature disentangling for video person re-identification. ICME .

Zhang, L., Shi, Z., Zhou, J.T., Cheng, M.M., Liu, Y., Bian, J.W., Zeng, Z., Shen, C., 2020a. Ordered or orderless: A revisit for video based person re-identification. IEEE TPAMI .

Zhang, T., Wei, L., Xie, L., Zhuang, Z., Zhang, Y., Li, B., Tian, Q., 2021b. Spatiotemporal transformer for video-based person re-identification. arXiv .

Zhang, W., Yu, X., He, X., 2017. Learning bidirectional temporal cues for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology .

Zhang, Z., Lan, C., Zeng, W., Chen, Z., 2020b. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Zhao, Y., Shen, X., Jin, Z., Lu, H., Hua, X.s., 2019. Attribute-driven feature disentangling and temporal aggregation for video person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.

Zheng, C., Wei, P., Zheng, N., 2021. A duplex spatiotemporal filtering network for video-based person re-identification, in: ICPR.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q., 2016a. Mars: A video benchmark for large-scale person re-identification, in: European Conference on Computer Vision.

Zheng, L., Yang, Y., Hauptmann, A.G., 2016b. Person re-identification: Past, present and future. arXiv .

Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T., 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in: IEEE / CVF Computer Vision and Pattern Recognition Conference.