Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.054895

REVIEW

Exploring Frontier Technologies in Video-Based Person Re-Identification: A Survey on Deep Learning Approach

Jiahe Wang[1], Xizhan Gao[1,*], Fa Zhu[2] and Xingchi Chen[3]

[1]Shandong Provincial Key Laboratory of Ubiquitous Intelligent Computing, School of Information Science and Engineering, University of Jinan, Jinan, 250022, China

[2]College of Information Science and Technology & College of Artificial Intelligence, Nanjing Forestry University, Nanjing, 210037, China

[3]Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen, 518066, China

*Corresponding Author: Xizhan Gao. Email: ise_gaoxz@ujn.edu.cn

Received: 11 June 2024 Accepted: 27 August 2024 Published: 15 October 2024

ABSTRACT

Video-based person re-identification (Re-ID), a subset of retrieval tasks, faces challenges like uncoordinated sample capturing, viewpoint variations, occlusions, cluttered backgrounds, and sequence uncertainties. Recent advancements in deep learning have significantly improved video-based person Re-ID, laying a solid foundation for further progress in the field. In order to enrich researchers' insights into the latest research findings and prospective developments, we offer an extensive overview and meticulous analysis of contemporary video-based person Re-ID methodologies, with a specific emphasis on network architecture design and loss function design. Firstly, we introduce methods based on network architecture design and loss function design from multiple perspectives, and analyzes the advantages and disadvantages of these methods. Furthermore, we provide a synthesis of prevalent datasets and key evaluation metrics utilized within this field to assist researchers in assessing methodological efficacy and establishing benchmarks for performance evaluation. Lastly, through a critical evaluation of the experimental outcomes derived from various methodologies across four prominent public datasets, we identify promising research avenues and offer valuable insights to steer future exploration and innovation in this vibrant and evolving field of video-based person Re-ID. This comprehensive analysis aims to equip researchers with the necessary knowledge and strategic foresight to navigate the complexities of video-based person Re-ID, fostering continued progress and breakthroughs in this challenging yet promising research domain.

KEYWORDS

Video-based person Re-ID; deep learning; survey of video Re-ID; loss function

1 Introduction

Person Re-ID has gradually come into our sight, with the advancement and development of science and technology. The earliest person Re-ID method was developed based on image data, which involve analyzing and comparing pedestrian images to determine the identity or Re-ID of individuals across different scenes. This technique has a wide range of applications in the fields of video

surveillance, intelligent transport, and security [1,2]. However, image-based person Re-ID methods face some challenges, such as pose change, view angle change, and occlusion in a single image, which limit its accuracy and robustness in real scenes. To solve these problems, video-based person Re-ID has gradually become the focus of research [3]. In intelligent video surveillance applications, video-based person Re-ID is defined as the identification of a single person from a large number of reference videos by various non-overlapping cameras [4]. Therefore, video-based person Re-ID can improve the accuracy of identifying by utilizing information about their movements in time series [5], and has gradually become a hot topic of attention and research.

Video-based person Re-ID extends image-based person Re-ID by incorporating temporal information from video sequences. The main difference lies in the input data, where video-based Re-ID handles video sequences while image-based Re-ID processes static images. Video sequences provide richer spatial-temporal information compared to static images. The key to video-based person Re-ID is to extract comprehensive pedestrian features from video sequences. This process follows similar principles as image-based methods while benefiting from the additional advantage of capturing richer spatial-temporal information. However, the challenges brought by inter-shot variations such as background clutter, occlusion, point-of-view, lighting changes, human posture changes, etc., are currently the main problems faced in the field of video-based person Re-ID [6]. In order to solve the problems, researchers have proposed many classical algorithms from the perspective of traditional machine learning and deep learning.

Traditional video-based person Re-ID methods were usually developed based on handcrafted features for distance metric learning between videos. They mainly focus on color features and less on texture features. For example, Farenzena et al. [7] proposed an appearance-based person Re-ID method that focuses on extracting features that capture three complementary aspects of human appearance: overall chromatic content, spatial arrangement of colors in stable regions, and features with high entropy. This approach enhanced the model's ability to cope with very low resolution, occlusion, and changes in pose illumination and viewing angle. On the other hand, Almasawa et al. [8] proposed a spatial-temporal segmentation algorithm to perform person Re-ID task. In this algorithm, HS histograms and edge histograms are computed for local regions, resulting in the generation of salient edge points. These salient edge points are robust to changes in the appearance of garments. The two features are organically combined to generate a feature that represents the entire pedestrian appearance. In Gray et al. [9], pedestrians were horizontally segmented using eight different color channels and 21 texture filters on the luminance channel. There are also many works [10–12] referred to the Gray et al. [9] with the rise of deep learning, the limitations of traditional video-based person Re-ID methods are gradually reflected, and the work on traditional methods is gradually decreasing.

Video-based deep person Re-ID methods are usually studied in terms of network architecture design and loss function design. In terms of network architecture design, researchers designed many networks to extract features from video frames and used feature fusion techniques to fuse these frame-level features to obtain a more robust pedestrian representation. In order to improve the differentiation of person representations, researchers have also adopted some regularization techniques such as data augmentation, attention mechanisms, etc., [13–15]. In terms of loss function design, researchers have designed some new loss functions for the characteristics of video-based person Re-ID. For example, considering the continuous action information of pedestrians in the video, some studies proposed a loss function based on triplet and multi-objective optimization to exploit the temporal information in the video [16]. In order to solve the occlusion problem, some studies have proposed loss functions based on the labelling of occluded regions.

For ease of understanding, we have compiled a timeline of the development of video-based person Re-ID tasks from 2014–2023 (as shown in Fig. 1), i.e., TDL [17] and other methods [18–21] are published between 2014 and 2016, DDCT [22], QAN [23], RQEN [24] and other methods [25–28] between 2017 and 2019, MGH [29] and other methods [30–33] in 2020, BiCnet-TKS [34], GRL [35] and other methods [36–40] in 2021,COSAM [41] and other methods [42–45] in 2022, DDCT [46], DSANet [47], IRF [48] and other methods [49–53] after 2023. These methods will be categorized and analyzed in the following text. Besides, the synthesized application scenarios, relevant contributions, and special considerations of previous survey papers are shown in Table 1.

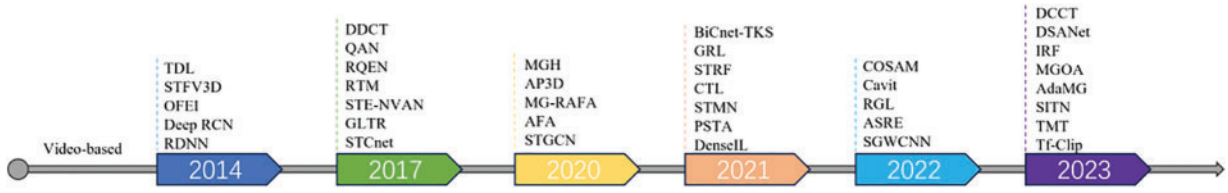

Figure 1: Video-based person re-identification task methodology timeline

Table 1: Comparison with other review papers

Survey	Covering	Analysis
Zheng et al. [1]	Image and video	Explored the history and relationships in person Re-ID. Hand-crafted and deep learning methods are comprehensively reviewed.
Ye et al. [15]	Image and video	Explored closed-world and open-world Re-ID. Baseline for single-/cross-modality Re-ID.
Mazzon et al. [54]	Crowd	Disadvantages of existing methods. Some improvements to the baseline.
Satta [55]	Appearance learning	Addresses current evaluation using public datasets. Introduced open and closed set Re-ID scenarios.
Bedagkar-Gala et al. [56]	Open-closed set	Raised open and closed set Re-ID scenarios. Highlighted public datasets with contemporary evaluations.
Lavi et al. [57]	Re-ID	Survey on techniques of deep neural networks. Addresses loss functions and data augmentation.
Wang et al. [58]	Re-ID	Traditional approaches and architectural perspectives. CNN, RNN and GAN for person Re-ID.

(Continued)

Table 1 (continued)

Survey	Covering	Analysis
Wu et al. [59]	Image	Various findings on Res-Net and inception architectures. A survey on deep-learning methods.
Masson et al. [60]	Image	Thoroughly examined pruning methods and strategies. Performance evaluation across diverse datasets.
Wang et al. [61]	Image and video	In-depth review of preceding Re-ID methodologies. Briefly discussed CNN, RNN and GAN.
Lin et al. [62]	Text and image	Thoroughly examined methods for person search. Feature learning and identity-driven methodologies.
Lin et al. [63]	Image and video	Extensively covered unsupervised methods. Discussion about dataset and evaluation. Performance analysis and metrics.
Ours	Video and loss function	Briefly discuss the methods in video Re-ID. Explored the loss function design. Performance analysis of current methods.

In summary, the contributions of this survey are summarized as follows:

(1) We propose to classify existing video-based deep person Re-ID methods into two categories according to their motivations, including methods for network architecture design and loss function design.

(2) Starting from the network architecture design, the researches on video-based deep person Re-ID in recent years are divided into seven different categories, and each category is discussed in some detail.

(3) Starting from the idea of loss function design, the related works are discussed in depth and divided into four categories.

(4) A brief overview of the commonly used datasets and evaluation criteria in the task of video-based person Re-ID is given, and finally some outlooks on the future development of the field are given.

2 Network Architecture

According to the network architecture, existing video-based deep person Re-ID methods can be further classified into the following seven categories:

(A) Global feature methods [64]: This type of method emphasizes the use of global attributes such as appearance and gait style for identification, that is, employs deep neural networks to capture

the overall features of pedestrian videos. (B) Local feature methods [65]: This type of method utilizes specific or local attributes in pedestrian videos, such as facial features or limbs, for identification. Common local feature learning methods include human keypoint localization and region segmentation. (C) Attention methods [66]: This type of method enhances Re-ID performance by focusing on crucial pedestrian areas in videos. The attention mechanism enables models to concentrate on key pedestrian parts like faces and hands for improved feature extraction. (D) Graph methods [67]: This type of method treats pedestrians in videos as graph nodes, extracting pedestrian features by establishing relationships between individuals. Graph methods can capture dynamic behaviors and interactions, offering richer information for identification. (E) Hypergraph methods: Unlike graphs, hypergraph models have more complex relationships among data points, facilitating a richer representation of inter-frame dependencies and enhancing the discriminative power of extracted features for accurate identification across video sequences. (F) Transformer methods [68]: This type of method utilizes self-attention mechanisms and positional encoding to capture temporal information and pedestrian interactions in videos, enhancing the understanding of dynamic video content and thus improving the accuracy of person Re-ID. (G) Generative methods: Through adversarial learning, this type of method can generate high-quality data for model training. It can also synthesize functional yet anonymized data, addressing privacy concerns while aiding in training.

Recent advances in the aforementioned seven approaches are detailed in the following section.

2.1 Global Feature Methods

This type of approaches extracts a single feature vector from each frame of a video individually, without relying on supplementary information. Finally, it clusters the features from each frame and synthesizes the video-level features (as shown in Fig. 2). Since deep neural networks were initially applied to image classification, global features learning was the primary choice in the early stages of integrating advanced deep learning techniques into the field of person Re-ID. Substantial progress has been made in recent research regarding global feature approaches. Researchers have enhanced the accuracy of person Re-ID by refining feature extraction and matching algorithms. Some research endeavors have explored the utilization of deep neural networks to extract more efficient global features. These networks, which typically use structures such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), can automatically learn the appearance characteristics of pedestrians and achieve efficient feature extraction. Moreover, some studies have concentrated on addressing pose variations and pedestrian occlusions. By introducing techniques such as pose estimation and occlusion detection, the performance of person Re-ID can be further improved.

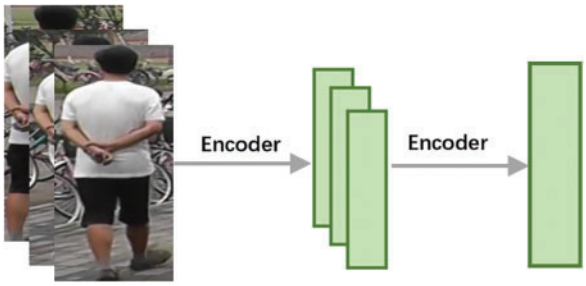

Figure 2: Schematic diagram of the global feature methods

When applying person Re-ID to the problem of human retrieval [69,70], researchers initially tended to focus more on matching local features and overlooked the significance of global features

due to its original purpose. With the introduction of deep learning methods into the field, the learning of global features was generally neglected in early research. As a pioneering work, the recurrent deep neural network (RDNN) [21] employed a pooling operation to initially the features of each frame. Subsequently, based on this, it utilized feature re-aggregation to aggregate all the time-step data into features for the entire video sequence using a feature reaggregation method. This innovative approach not only simplified the data processing process, but also notably enhanced the accuracy of person Re-ID.

For an in-depth comparison of different temporal modeling approaches, Gao et al. [25] conducted a comprehensive study of 3D Conv NET, RNN, temporal pooling, and temporal attention models using cross-entropy, and triplet loss. These methods find wide applications in person Re-ID, making their comparative analysis crucial for understanding the nature of person Re-ID and enhancing performance. A novel spatial-temporal attention was employed in the STA [71] to achieve frame-level feature aggregation. This method deviated from the conventional average pooling technique by generating a two-dimensional attention score matrix through inter-frame regularization. This matrix facilitated the distinction of spatial part importance across different frames, effectively handling scenarios involving pose changes and occlusion in video-based person Re-ID. On a different note, Liu et al. [72] proposed a memory unit that recovers missing parts and suppressed noisy segments in the current frame features by referring to historical frames. They supervised the model training using a multilevel training strategy.

Recently, Li et al. [73] proposed multiple Hellinger distance-based regular terms to ensure that the model can learn all parts of the person's body simultaneously. The model can effectively overcome occlusions and misalignments, allowing for the extraction of more representative features. In addition, the Attribute-Driven Feature Disentangling and Frame Re-weighting [74] model has also been implemented in video-based person Re-ID. Initially, the features of a single frame were decomposed into multiple sub-feature groups, with each sub-feature corresponding to a specific semantic attribute. Subsequently, the sub-features were re-weighted based on the confidence level of attribute Re-ID. Finally, these weighted sub-features were aggregated along the time dimension to construct the final representation. This methodology enhanced the most informative regions in each frame, resulting in a more discriminative sequence representation. In order to better leverage short-long term temporal cues, Li et al. [27] proposed the GLTR method, a two-stream network model that enhanced convolution by incorporating short-term cues and capturing long-term relationships through a temporal self-attention model. This approach alleviated occlusion and noise in the video and aggregated the short-term and long-term cues into a final GLTR via a single-stream CNN.

The global-guided reciprocal learning (GRL) framework, as proposed in [35], extracted fine-grained information from image sequences. Based on local and global features, the Global Guided Correlation Estimation (GCE) module generated feature correlation maps for localizing low and high correlation regions to identify similar people. Moreover, to handle multiple memory units and enhance temporal features, the paper also introduced construct Temporal Reciprocal Learning (TRL) to collect specific cues. Yang et al. [43] proposed a novel Relation-Based Global-Partial Feature Learning (RGL) framework to explore discriminative spatiotemporal features by utilizing global and partial relationships between frames. Additionally, Chai et al. [44] proposed a novel network architecture named Attribute Salience Assisted Network (ASA-Net) for attribute-assisted video person Re-ID. Both of RGL and ASA-Net have demonstrated advanced performance.

In conclusion, global feature representation learning methods offer the advantages of extracting comprehensive information from pedestrian images, showcasing enhanced robustness and generalizability. The joint aggregation of frame-level features and spatial-temporal appearance information plays a pivotal role in video representation learning. Nonetheless, as global feature representation overlooks the local details in pedestrian images, there may be some limitations for individuals with similar global features but different local appearance. Consequently, some hybrid feature representation learning methods combining global features and local features have emerged in recent years to further enhance the accuracy and robustness of person Re-ID.

2.2 Local Feature Methods

Local feature learning entails the focused acquisition of pedestrian's local features, stemming from the understanding that distinct individuals harbor rich identity information within their specific body parts. Targeted learning aids in feature alignment, enhancing the model's robustness [75]. The segmentation of body parts is achieved through automated generation via human parsing (pose estimation) algorithms or rough horizontal segmentation (for details see Fig. 3).

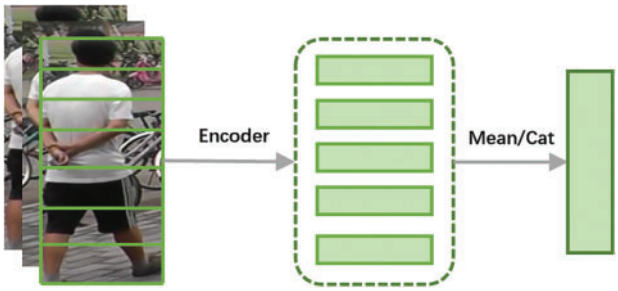

Figure 3: Schematic diagram of the local features approach

For automatic body part detection, a prevalent approach involved integrating whole-body representations with local features of specific parts. Techniques such as multi-channel aggregation, multi-scale context-aware convolution, multi-level feature decomposition, and bilinear pooling have been developed to enhance the learning of local features. Some studies have explored part-level similarity combination rather than feature-level fusion. Additionally, another common strategy was leveraging pose-driven matching, pose-guided partial attention modules, and semantic part alignment to bolster the model's resilience to background clutter.

In the domain of horizontally segmented region features, a Part-based Convolutional Baseline (PCB) [65] emerged as a robust baseline for learning part features and remains prominent in the current state-of-the-art. While pooling techniques reduced the number of parameters and introduced translation invariance, they can also destroy valuable structural relationship information. Bao et al. [76] proposed an approach Structural Relationship Learning (SRL), a method that captured structural relationships by creating a spatial structural graph based on convolutional features. This approach propagated information over edges, combines pooling operations with metric fusion, and significantly enhances model robustness. Another popular research was the Spatial-Temporal Complementary Network STC-Net [28], which addressed partial occlusions by explicitly reconstructing the appearance of obscured parts. Song et al. [24] proposed the Region-Based Quality Estimation Network (RQEN), a model that efficiently learned and integrated information from all frames within a sequence. By leveraging complementary details across frames, RQEN effectively compensated for lower-quality

image regions by incorporating superior regions from other frames. This helped improve the robustness of person Re-ID, particularly in scenarios involving complex video scenes with partial noise.

Unlike previous methods, Hou et al. [77] proposed a temporal complementary learning network to extract complementary features from continuous video frames for video person Re-ID. The network consisted of a temporal saliency erasure (TSE) module and a temporal saliency boosting (TSB) module. The TSE module mined the complementary features of consecutive video frames through saliency erasure operations and an ordered learner, gradually aggregating the local features to ultimately form the overall features of the target identity. To extract more detailed clues, Multi-Granularity Reference-Aided Attention Feature Aggregation (MG-RAFA) [31] was proposed. The module meticulously aggregated spatial-temporal features into differentiated video-level feature representations by learning attention from a global perspective through convolutional operations. The distribution of attention was inferred by overlaying the relationships of features with reference feature nodes representing global information. Additionally, Li et al. [78] took into account the overall visual similarity between video frames while focusing on improving the quality recovery of misaligned parts. And Leng et al. [49] proposed a novel video person Re-ID framework, called Multi-Granularity Occlusion Aware (MGOA), which extracted multi-granularity features by precisely erasing the occlusion.

2.3 Attention Methods

Due to the development of person Re-ID datasets and the increasing complexity of the applied scenarios, models that relied solely on global features or local features don't achieve satisfactory results. By using an attentional mechanism, the models are able to focus more flexibly on specific regions in the image, thereby increasing the sensitivity of the model to local information (as shown in Fig. 4). This compensatory mechanism helps to address the limitations of the local features approach when dealing with complex scenes or in the presence of occlusions, allowing the model to better capture critical local details and improve overall performance. Thus, with the introduction an attention mechanism, the models are able to compensate for some of the limitations of local features approaches, allowing them to handle image tasks more comprehensively and efficiently. The most typical attention mechanisms include pixel-level attention, channelized attention, and methods to suppress background information.

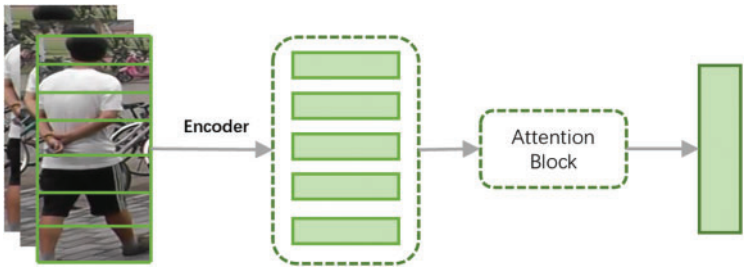

Figure 4: Schematic diagram of the attention methods

Song et al. [79] first introduced binary segmentation masks to create synthetic RGB mask pairs as inputs. Subsequently, they designed Mask-Guided Contrastive Attention Model (MGCAM) to separately learn features of the body and background regions. Additionally, they proposed a new region-level triplet loss to supervise the model training, pioneering the application of region-level contrast learning in person Re-ID. Similar to the previous task, to address the interference of cluttered background on the model, a novel co-segmentation inspired deep architecture was proposed, named

Co-Segmentation based Attention Module (COSAM) [80]. This module enhanced the model performance in an unsupervised manner by utilizing a set of co-significant features that consistently activated together across multiple frames of the video. Building upon previous work, Subramaniam et al. [41] suggested that the concept of co-segmentation enables the model to automatically focus on task-specific salient regions and enhance the performance of the underlying task in an end-to-end manner. These segmentation methods were capable of extracting unique features from character images and use them for channel and spatial attention. In another study, STAL [81] focused on spatial-temporal feature learning by segmenting the video into multiple spatial-temporal units. They established a scoring mechanism to assess each unit, with the ultimate objective of ensuring that the model could extract more valuable information.

In real-world applications, human movement patterns serve as crucial cues for Re-ID. A flow-guided mutual attention network can effectively leverage this information [82]. This model can seamlessly integrate both image and optical flow sequences to enhance the encoding of spatial-temporal information. It also introduced a method to aggregate features for longer input streams to enhance the representation of video sequences. Recent studies have concentrated on multi-grained and multi-attention approaches to make models more attentive to essential parts of the human body. It has been shown that occlusion is still a great challenge in person Re-ID, and the attention mechanism can effectively address this issue. Consequently, the Concentrated Multi-grained Multi-attention Network (CMMA-Net) [83] was introduced. This network incorporated two attention modules to fully exploit the information in the multi-scale feature maps and introduced a diversity loss function to ensure that the model can focus on multiple regions. Tao et al. [48] presented an adaptive Interference Removal Framework (IRF) to learn discriminative feature representations by eliminating various interferences. And Zhao et al. [84] proposed a novel Multi-scale Feature Aggregation Network (MFANet) to solve the challenge of occlusion and misalignment in the video. Hou et al. [34] proposed the Bilateral Complementary Network (BiC-Net), a two-branch network where one processes raw resolution images to capture detailed visual cues, while the other branch uses down-sampling to extract contextual information. This design enabled the model to focus on different body parts in consecutive frames, facilitating a more comprehensive capture of overall features. In contrast to prior studies, Chen et al. [3] didn't segregate spatial and temporal information for subsequent aggregation. Instead, they employed an end-to-end 3D framework to concurrently capture spatial-temporal information, leveraging both temporal and spatial aspects. By utilizing the joint constraints of temporal and spatial attention, the model's robustness was enhanced. Bai et al. [85] introduced the Salient-to-Broad Module (SBM) and Integration-and-Distribution Module (IDM) within the Attention Mechanisms. These modules aimed to alleviate the limitations of conventional attention mechanisms, which often excessively focus on image localization and similarity. The goal was to enhance model performance, particularly in video-based person re-identification tasks. On a different note, Kim et al. [47] proposed a feature decoupling model that separates view angle information from pedestrian identity information. This separation allowed the model to better handle challenges like view angle variations and changes in illumination, thus bolstering the model's robustness. This innovation introduced a novel approach to enhancing performance in video-based person Re-ID.

There are still certain applications that require the use of large models. For instance, Yu et al. [53] proposed a novel one-stage Text-Free CLIP-based learning framework named TF-CLIP for video-based person Re-ID. The utilization of CLIP in video-based person Re-ID offers several advantages. Firstly, CLIP demonstrates strong visual and semantic representation capabilities, capturing intricate associations between images and text. Secondly, CLIP is pre-trained on large-scale datasets, enabling it to generalize effectively and adapt to various challenges encountered in video Re-ID, such as

viewpoint variations, object occlusions, and cluttered backgrounds. This enhancement contributes to the accuracy and robustness of person Re-ID.

2.4 Graph Methods

With the remarkable success of convolutional neural networks (CNNs) [86] in the field of image understanding and reconstruction, researchers in academia and industry have begun to focus on the development of convolutional methods for graph-structured data. In the domain of video-based person Re-ID, the utilization of graph methods [87] has garnered significant attention (as shown in Fig. 5).

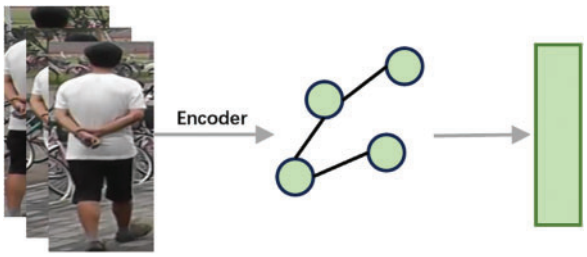

Figure 5: Schematic diagram of the graph methods (each frame of the image is viewed as a node, and each line connected with two nodes)

Yan et al. [88] were pioneers in combining video-based person Re-ID with graph model. Their paper proposed the Recurrent Feature Aggregation Network (RFA-Net), which enhanced the aggregation of pedestrian features at each timestamp to generate highly discriminative video-level pedestrian feature representations. Inspired by ternary and contrast loss, Cheng et al. [89] proposed a structured Laplacian embedding algorithm. This algorithm combines the concept of deep metric learning with graph model structures to leverage the relationships between samples. By exploiting the structured distance relationship between samples, this integration enhances the model's performance. In Reference [67], a new unsupervised model was proposed, establishing relational information between different probe-graph library samples. This advancement in accuracy of similarity estimation for challenging samples resulted in optimized ranking mechanism in video-based person Re-ID. The primary objective of video-based person Re-ID is to accurately calculate the visual similarity between person images. However, many existing models calculate the similarity between different gallery and query samples individually, neglecting the knowledge between each pair of gallery samples or query samples. Consequently, researchers have started exploring methods to effectively harness the potential associations between different samples, a focal point in numerous studies.

To solve the above problem, Chen et al. [90] introduced a local similarity metric that maximizes image correlation within each video frame. It incorporates an approximate inference scheme to estimate video-to-video similarity. Crucially, the model was trained in an end-to-end manner, ensuring comprehensive and seamless learning. To tackle challenges like varying lighting, large pose changes and pedestrian occlusion, the contextual information between neighboring frames was used to assist the Re-ID. Yan et al. [91] introduced a context instance expansion module, altering the model to take each frame's feature map as input and employ the attention mechanism to extract useful contextual information. Additionally, a graph learning framework was constructed to enhance the model's discrimination capabilities.

To enhance the utilization of temporal and spatial attention to extract better local features, Wu et al. [92] proposed an innovative adaptive graph representation scheme designed to leverage correlation information among local features. A critical aspect of this approach involved constructing an adaptive structure-aware neighborhood graph. This graph was created by utilizing pose-aligned connections and feature affine connections to effectively model intrinsic relationships between graph nodes. As a result, more refined local features were acquired and subsequently fused into global features. This fusion process bolstered the similarity between similar samples, thereby enhancing performance in handling homogeneous samples. Liu et al. [37] partitioned the CNN-extracted feature graph into multi-scale features and utilized individual local features from these multi-scale features as graph nodes. Consequently, they constructed a topology based on contextual enhancement. Additionally, they employed 3D convolution to facilitate the simultaneous extraction of spatial-temporal features at various scales, enabling better mining of spatial-temporal cues complementary to appearance information and improving the model's performance. In contrast, Chen et al. [93] adopted a different approach by extracting features from key points of the human body and constructing a graph model using these key points. The features of these key points were then propagated through the connected nodes, allowing for message passing and updates. More recently, Yao et al. [45] proposed a novel Sparse Graph Wavelet Convolution Neural Network (SGWCNN) to effectively solve the problems of short time occlusion and pedestrian misalignment. This method aimed to enhance the performance of human pose modeling in person Re-ID.

2.5 Hypergraph Methods

In contrast to graph structures, hypergraphs permit a single edge to connect multiple nodes, known as hyperedges, whereas graphs usually have only binary relationships, with each edge connecting two nodes. The hypergraph structure offers greater flexible and suitable for describing complex multivariate relationships (as shown in Fig. 6). Over recent years, hypergraph models have been used in person Re-ID, action recognition [94] and image recognition [95].

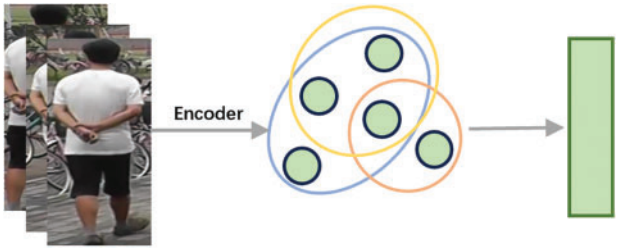

Figure 6: Schematic diagram of the hypergraph methods (each edge is connected with more than two notes)

Hypergraph methods are specifically designed to handle the unique structure of hypergraphs. As the hyperedges in hypergraphs can link multiple nodes, hypergraph methods perform well in managing and analyzing complex multivariate relationships between nodes. The commonly used hypergraph methods include hypergraph matching, hypergraph cut, and hypergraph embedding. Hypergraph matching algorithms are instrumental in identifying patterns within hypergraphs, while hypergraph cut algorithms facilitate partitioning and segmentation of hypergraphs. In contrast, hypergraph embedding algorithms focus on embedding hypergraphs into lower-dimensional spaces to enhance structural analysis. At present, hypergraph methods have been applied in the field of video pedestrian Re-ID. For example, Yan et al. [29] introduced a Multi-Granular Hypergraph

(MGH) model, which adeptly captures multi-granular spatial-temporal dependencies by constructing hypergraphs with varying spatial granularities using part-based features in video sequences.

2.6 Transformer Methods

Transformer models have demonstrated their effectiveness in capturing complex temporal dynamics in videos compared to traditional approaches. Unlike traditional methods that rely on handcrafted features or using recurrent neural networks (RNNs) for sequential data processing. Transformer models leverage self-attention mechanisms to capture long-range dependencies and global context information. By focusing on relevant temporal relationships across the entire sequence, they excel in modeling complex dynamics on person Re-ID scenarios. As shown in Fig. 7, they can effectively handle variations in pose, viewpoint, and appearance over time, making them particularly suitable for video-based person Re-ID.

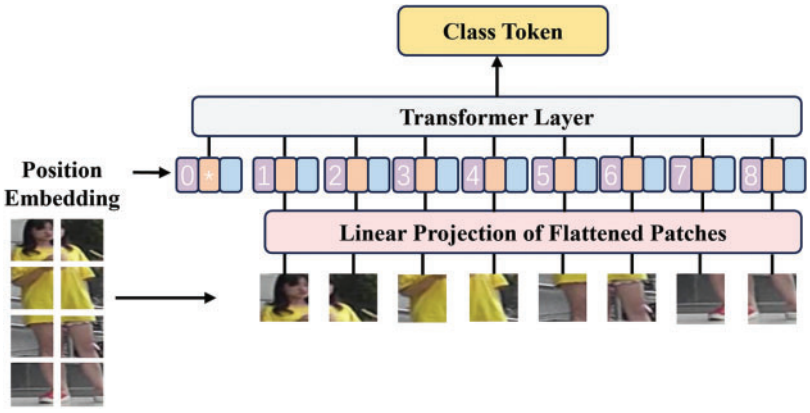

Figure 7: Schematic diagram of the Transformer methods

Inspired by recent research advances, Zhang et al. [96] proposed the Spatiotemporal Transformer (STT), marking its inaugural application in video-based person Re-ID. The model effectively addressed the overfitting issue and yielded crucial insights for subsequent research endeavors. In a similar vein, Liu et al. [52] proposed the Trigeminal Transformers (TMT) module, which decoupled raw video features into three different forms: temporal, spatial, and spatial-temporal domains. This module extracted multi-view features from these three forms to capture fine-grained features and amalgamate them into multi-view representations. To bolster single-view features, self-view converters were introduced, while cross-view converters were employed to combine multiple features. Notably, they presented the Duplex Spatial Temporal Filtering Network (DSFN) [97], a dual-stream network designed to focus on separately extracting static and dynamic features from each sample. Unlike conventional filters, this model emphasized the extraction of fine-grained features. Furthermore, the incorporation of a sparse orthogonal constraint matrix in this work compelled the model to prioritize local features in diverse regions, thereby enhancing the discriminative capacity of the features.

Some models use both convolutional and transformer models simultaneously. For example, Liu et al. [46] proposed a novel spatial-temporal complementary learning framework named Deeply-Coupled Convolution-Transformer (DCCT). This approach mechanically integrates two types of typical features from the same videos to create more informative representations. He et al. [40] proposed a learning framework called Dense Interaction Learning (Dense-IL), which combined a CNN and an attention-based architecture to address challenges in video-based person Re-ID.

The framework comprises a CNN encoder for extracting spatial features and a dense interaction decoder for densely modeling cross-frame spatial-temporal interactions. It emphasizes moderately fine-grained CNN features to generate multi-granular spatial-temporal representations. Additionally, the CAViT model proposed in [42] tackled spatial-temporal alignment issues present in traditional models by fully leveraging potential interaction information between spatial-temporal features. It extracted more accurate representations of the features. Xu et al. [22] maximized the potential of both convolutional neural networks and transformer models by introducing the Deeply-Coupled Convolution-Transformer (DCCT) model. This model adeptly integrates features extracted from the two different architectures, enhancing overall performance and laying a solid foundation for subsequent dual-stream network research endeavors.

2.7 Generative Models

The generative models mainly include Generative Adversarial Networks (GAN), and Variational Autoencoders (VAE). GAN is composed of a generator network and a discriminator network trained in an adversarial manner. The generator learns to synthesize realistic samples, while the discriminator aims to distinguish between real and generated samples. This adversarial training process leads to the generation of high-quality, realistic data. In contrast, VAEs are probabilistic models that learn to encode and decode data. VAE network aims to find a latent representation that captures the underlying structure of the data. By sampling from this latent space, VAE network can generate new samples that closely resemble the training data. The generative models have been widely used in image generation, text modeling, and anomaly detection tasks. The utility of GAN and VAE extends beyond data generation, they can be leveraged to synthesize anonymized yet functional data for training, addressing privacy concerns.

Recently, many effective generative models have been proposed. For instance, Ge et al. [98] proposed a solution that involved training an image-to-image translation network using the Cycle-GAN architecture. Reference [50] introduced an Adaptive Memory with Group Labeling (AdaMG) framework for unsupervised person Re-ID. This framework aimed to improve the performance of the model by incorporating adaptive memory mechanisms and group labeling techniques. Their approach centered on bidirectionally translating images using appropriate generators, effectively bridging the domain gap. Additionally, diversifying the training data is crucial to bolster the model's generalization capability. To address the impact of perspective changes, Wu et al. [99] proposed approach introduces a novel few-shot deep learning method for video-based person Re-ID. It leveraged variational recurrent neural networks (VRNNs) and adversarial training to learn discriminative and view-invariant representations. The method focused on producing latent variables with temporal dependencies that exhibit high discriminability while being invariant to viewpoint variations when matching individuals.

The generative models also can be used in person Re-ID to generate synthetic pedestrian images that closely resemble real images while protecting privacy. By employing these generative models, training datasets can be enriched with non-sensitive synthetic images, enabling the development of robust person Re-ID models. For example, in Reference [100], a GAN-based approach was proposed to generate synthetic images for person Re-ID while preserving identity information, demonstrating the potential of generative models in enhancing privacy protection in this domain.

3 Loss Function

Loss function plays an important role in discriminative feature learning. Usually, the cross-entropy loss separates the learned features instead of discriminating them, which plays a limited role in person Re-ID, so many researchers instead started with the idea of optimizing the loss function and also made a lot of progress [101]. This paper focuses on describing the loss functions used to supervise deep neural networks, which are mainly categorized into four types, i.e., identity loss, verification loss, triplet loss and OIM loss.

3.1 Identity Loss

Some method viewed the training process of video-based person Re-ID model as an image classification problem, e.g., each identity is a class. However, in the testing phase, the output of the pooling or embedding layer was used as the feature vector. Given an input video x_i labeled as y_i, the predicted probability of x_i being recognized as class y_i was encoded in a soft-max function denoted as $p(y_i|x_i)$.

$$L_{id} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(y_i|x_i\right)\right) \tag{1}$$

where n denotes the number of training samples in each batch. Identity loss is widely used in current models [102], it can almost be said that identity loss is indispensable whenever supervised deep learning models are used. There are many reasons for its widespread use, first of all, this loss function is very easy to use, requiring only a simple linear classifier, and converges faster. Some improvements have been made in other works [103], such as in References [104] and [105]. There are also some methods [106,107] used labeled smoothing, a simple but effective strategy, to prevent overfitting and improve the generalization ability of the model [108].

3.2 Verification Loss

Verification loss is a loss function used in deep learning to handle the verification task, which is usually to test whether two belong to the same class, by comparison loss [109] or binary verification loss [110,111] to optimize the relationship between a pair of similar or dissimilar samples.

$$L_{con} = \left(1-\delta_{ij}\right)\left\{\max\left(0,\rho-d_{ij}\right)\right\}^2 + \delta_{ij}d_{ij}^2 \tag{2}$$

where d_{ij} denotes the distance between the feature vectors x_i and x_j of the two input samples. δ_{ij} is a control symbol ($\delta_{ij}=1$ when x_i and x_j belong to the same unit, otherwise $\delta_{ij}=0$). ρ is a hyperparameter.

Some binary validation work aims at distinguishing whether the input image pairs belong to the same class or different classes, so as to have the effect of supervised model training. Typically, we use $f_{ij}=(f_i-f_j)^2$ to denote the difference metric between two different samples, where f_i and f_j are feature vectors from two different samples x_i and x_j. We then use this metric to judge whether these two features are positive (positive pair) or negative (negative pair). We use $p(\delta_{ij}|f_{ij})$ to denote the probability that the input sample pair (x_i, x_j) is labeled as δ_{ij} (0 or 1), which in combination with the cross-entropy allows us to write the loss function as follows:

$$L_{veri}\left(i,j\right) = -\delta_{ij}\log\left(p\left(\delta_{ij}|f_{ij}\right)\right) - \left(1-\delta_{ij}\right)\log\left(1-p\left(\delta_{ij}|f_{ij}\right)\right) \tag{3}$$

If this loss function is only used to supervise the training of the model, it may not fully utilize the identity information of each sample in the training set, and thus is usually used in conjunction with identity loss [112].

3.3 Triplet Loss

The method treats the training process of the Re-ID model as a retrieval sorting problem. The basic idea is that the distance between pairs of samples of the same identity should be smaller than the distance between pairs of samples of different identities, and this distance difference should be smaller than a predefined boundary. Typically, a triad contains an anchor, a positive sample (of the same identity as the anchor) and a negative sample (of a different identity than the anchor). The triplet loss with margin parameter is denoted as:

$$L_{tri}(i,\ j,\ k) = \max\left(m + d_{ij} - d_{ik},\ 0\right) \tag{4}$$

where d_{ij} represents the Euclidean distance between positive samples, d_{ik} represents the Euclidean distance between negative samples, and m represents the margin. If the above loss function is optimized directly, a significant proportion of simple triples will dominate the training process, and some of these low-quality pairs of triples will be useful for misleading the training of the model, thus limiting the discriminative ability of the model. To alleviate this problem, researchers have designed various ternary mining methods [113,114]. The basic idea is to select informative triples [113,115]. Specifically, a modest positive mining with weight constraints is introduced in Shi et al. [115], which directly optimizes the feature differences. Hermans et al. [113] demonstrated that not every pair of triplets plays a positive role during training, and thus it is beneficial to perform the hardest positive and negative mining online in each training batch to enhance the discriminative properties of the Re-ID model. Meanwhile, some approaches investigate point-to-set similarity strategies for informative triad mining [116,117], enhances the robustness to outlier samples through hard and soft mining schemes. Furthermore, Jiang et al. [118] proposed Weighted Triple-Sequence Loss (WTSL) based on the traditional triple-sequence loss, which is a function that helps to reduce the frame-based image-level information by explicitly encoded as video-level features, which helps to reduce the effect of outlier frames. In WTSL, similar videos are made closer by adjusting the intraclass distance, while increasing the interclass distance to make different videos farther apart. In addition, some works such as [119], have used a similar approach to improve the traditional triplet loss to better incorporate the video-based person Re-ID.

To further enrich the learnable information in each set of samples, Chen et al. [120] proposed a deep network of quaternions, where each quaternion contains an anchor sample, a positive sample and two mined negative samples, which contain richer potential information compared to ternary groups. The formation of the quaternion is achieved by online hard negative mining based on margins. By optimizing these four correlation samples, intra-class variation can be reduced and inter-class variation can be increased.

The combination of triplet loss and identity loss is one of the most popular solutions for deep Re-ID model learning [121–123]. These two components are mutually beneficial in discriminative feature representation learning, with the former aiming to bring similarity closer to pull dissimilarity apart, and the latter aiming to learn essential identity information in the sample.

3.4 OIM Loss

In addition to the above three loss functions, Xiao et al. [124] designed Online Instance Matching loss (OIM), which additionally designs a memory module and utilizes a correlation mining algorithm to uncover sample pairs in the batch that are more favorable for model training. A memory module

$\{e_k,\ k = 1,\ 2,\ \cdots, c\}$ where the features of different categories are stored and c denotes the category number. the OIM loss is:

$$L_{oim} = -\frac{1}{n}\sum_{n=1}^{n} log\frac{\exp\left(\frac{v_i^T f_i}{\tau}\right)}{\sum_{k=1}^{c}\exp\left(\frac{v_k^T f_i}{\tau}\right)} \tag{5}$$

where v_i denotes the features of category y_i stored in the memory module, and τ is the temperature parameter controlling the feature embedding space. $v_i^T f_i$ is used to calculate this sample's online instance matching score for this sample, so that the impact of mislabeled samples in the dataset on model training can be well avoided. This memory scheme is also used in unsupervised domain adaptive Re-ID [125]. Furthermore, Pathak et al. [126] introduced CL centers online soft-mining loss, by using the center vector of the center loss as the class label vector representation, frames that contain higher noise can be excluded because the center vector has higher variance compared to the original classifier weights.

4 Datasets and Evaluation Indicators

4.1 Datasets

Currently, the commonly used public datasets in the field of video-based person Re-ID include: PRID-2011, MARS, Duke MTMC-Video, iLIDS-VID.

PRID-2011 [127]: PRID-2011 is a dataset collected by the Austrian Institute of Technology (AIT) in which video pairs are recorded from two stationary surveillance cameras. Camera A contains 385 pedestrians and camera B contains 749 pedestrians, but only 200 pedestrians appear in both cameras. These images show differences due to factors such as variations in camera viewpoints, illumination differences, pedestrian backgrounds, and camera characteristics. Since the images are extracted from trajectories, each person has multiple different poses in the images. This dataset makes an important contribution to the study of person Re-ID.

MARS (Multi-Camera Activity Dataset) [128]: MARS is a large-scale multi-camera activity dataset that was created by researchers at The Chinese University of Hong Kong. The dataset captures video sequences taken from six different camera views, simulating the setup of a real surveillance system. The dataset contains 1261 pedestrian identities and includes more than 20,000 video clips. The pedestrians in these video clips travel between different cameras, showing diverse behaviors and perspective changes. Compared to other datasets, MARS is more challenging because it more realistically reflects the complex scenarios and multi-camera cross-views found in actual surveillance systems.

Duke MTMC-Video [102]: Duke MTMC-Video Re-ID is a video-based person Re-ID dataset extended from the Duke MTMC dataset, which was created by researchers at Duke University (Duke). It contains 2196 pedestrian identities and 2535 video sequences from 8 cameras. The dataset presents a rich variety of viewpoints and environments, and is used to evaluate the robustness and generalization ability of the model for person Re-ID.

iLIDS-VID [6]: iLIDS-VID is a small video-based person Re-ID dataset that was created by the University of Southampton (UK) and contains 300 pedestrian identities and 600 video sequences. The dataset provides pedestrian videos from two cameras, which helps to evaluate the performance of the model in more complex environments and serves as a catalyst in the field of video-based person Re-ID. The specifics of the four datasets are shown in Table 2.

Table 2: Comparison of video-based person re-identification datasets

Dataset	Time	ID	Track	Cam	Label
PRID-2011	2011	200	400	2	Hand
iLIDS-VID	2014	300	600	2	Hand
MARS	2016	1261	20715	6	Auto
Duke MTMC-Video	2018	1812	4832	8	Auto

4.2 Evaluation Indicators

The most commonly used evaluation indicators include cumulative matching characteristic curve, mean average precision, and Top-k accuracy.

Cumulative Matching Characteristic (CMC) Curve: the CMC curve is one of the most important metrics for evaluating the performance of person Re-ID. It measures the ranking of recognition algorithms that are able to find the identity of pedestrians at different matching accuracies. By plotting the CMC curve, it is possible to visually compare the performance differences between different methods on different rankings.

Mean Average Precision (mAP): mAP is another commonly used evaluation metric that combines the accuracy and recall of the retrieval results. mAP calculates the average accuracy under different thresholds and measures the performance of the algorithms under different recall rates.

Top-k Accuracy: In addition to CMC curve and mAP, Top-k accuracy is also commonly used to evaluate the performance of person Re-ID algorithms. It indicates whether the correct identity is included in the Top-k most similar candidates.

5 Analysis and Outlook

5.1 Performance Analysis of Existing Methods

We list in Table 3 the results of the evaluation of each model on the four main datasets in recent years and indicate the baseline used for each modeling approach.

Table 3: Performance of different methods on MARS, Duke-V, iLIDS-VID and PRID-2011 datasets

Method	Backbone	MARS (Rank-1)	Duke-V (Rank-1)	iLIDS-VID (Rank-1)	PRID-2011 (Rank-1)
TDL [17]	Resnet-50	×	×	56.3	56.7
STFV3D [18]	Resnet-50	×	×	44.3	64.1
OFEI [19]	Resnet-50	×	×	69.1	66.7
Deep RCN [20]	Resnet-50	×	×	42.6	49.8
QAN [23]	Resnet-50	×	×	68.0	90.3
STE-NVAN [26]	Resnet-50	88.9	95.2	×	×
GLTR [27]	Resnet-50	87.0	96.3	86.0	95.5
VRSTC [28]	Resnet-50	88.5	95.0	83.4	×
MGH [29]	Resnet-50	90.0	×	85.6	94.8

(Continued)

Table 3 (continued)

Method	Backbone	MARS (Rank-1)	Duke-V (Rank-1)	iLIDS-VID (Rank-1)	PRID-2011 (Rank-1)
AP3D [30]	AP3D	90.1	96.3	86.7	×
MG-RAFA [31]	Resnet-50	88.8	×	88.6	95.9
AFA [32]	Resnet-50	90.2	97.2	88.5	×
STGCN [33]	Resnet-50	90.0	97.3	×	×
GRL [35]	Resnet-50	91.0	×	90.4	96.2
STRF [36]	Resnet-50	90.3	97.4	89.3	×
STMN [38]	Resnet-50	89.9	96.7	80.6	×
PSTA [39]	Resnet-50	91.5	98.3	91.5	95.6
Dense-IL [40]	Resnet-50	90.8	97.6	92.0	×
Cavit [42]	ViT	90.8	×	93.3	95.5
RGL [43]	Resnet-50	89.1	97.2	88.7	×
ASRE [44]	Resnet-50	90.6	97.6	×	×
SGWCNN [45]	Resnet-50	90.0	96.3	87.8	×
DDCT [46]	Resnet-50	92.3	**98.4**	91.7	**96.8**
DSA-Net [47]	Resnet-50	91.1	97.2	85.1	×
SITN [51]	ViT	89.2	96.0	94.0	96.6
TMT [52]	Resnet-50	91.2	×	91.3	96.4
SSP [53]	Clip	**93.0**	×	**94.5**	×
STA [71]	Resnet-50	86.3	96.2	×	×
STAN [73]	Resnet-50	82.3	×	80.2	93.2
ADFD [74]	Resnet-50	87.0	×	86.3	93.9
TCL-Net [77]	Resnet-50	89.8	96.9	86.6	×
SBM [85]	Resnet-50	91.0	×	92.5	96.5
Snippet [129]	Resnet-50	86.3	×	85.4	93.0
HMN [130]	Resnet-50	89	96.2	×	×
SANet [131]	Resnet-50	91.2	97.7	×	95.5
DPRAM [132]	Resnet-50	89	97.1	×	×

Firstly, we can observe that different methods and models achieved varying rankings on different datasets. For the MARS dataset, the SSP [48] method achieved the top rank with a Rank-1 accuracy of 93.0%, followed by the DDCT [46] method and the PSTA [39] method with Rank-1 accuracies of 92.3% and 91.5%, respectively. On the Duke-V dataset, the DDCT [46] method again performed well, achieving a Rank-1 accuracy of 98.4%, closely followed by the PSTA [39] method with Rank-1 accuracies of 98.3%. On the iLIDS-VID dataset, the SSP [48] method secured the top rank with a Rank-1 accuracy of 94.5%, followed by the SBM [85] method and the Cavit [42] method with Rank-1 accuracies of 92.5% and 93.3%, respectively. Finally, on the PRID-2011 dataset, the DDCT [46] method achieved a Rank-1 accuracy of 96.8%, claiming the first position, followed by the SITN [51] method and the SBM [85] method with Rank-1 accuracies of 96.6% and 96.5%, respectively.

Secondly, we can observe that the choice of backbone network also influenced the results. While most methods utilized the Resnet-50 backbone, some methods opted for other networks, such as

the AP3D [30] method using the AP3D network and the Cavit [42] method using the ViT network. This suggests that selecting an appropriate backbone network is crucial for the performance of person Re-ID.

5.2 Outlook of Future Works

The design ideas of most good models are basically similar, i.e., a spatial learning module is first designed to better extract the features of each frame in the video, and then a temporal information aggregation module is designed to aggregate this spatial information into the final video features. However, in real-world application scenarios, the description of visual information contains many different forms, e.g., identity labeling, camera labeling, etc. Most studies have matched probe images to gallery maps through visual similarity while ignoring more valuable information, such as textual information. The information richness of a text we use to describe an image or a video is much more than an identity tag, and the cross-modal and multi-modal tasks have received more and more attention from scholars in terms of the number of published articles on person Re-ID.

At the same time, accurate labeling of new datasets takes a lot of time and effort. In many cases, labeled data is prone to errors due to a variety of factors such as person visibility, background clutter, and noise in the image. Some researchers have focused on unsupervised methods [133] and active learning methods [134] to alleviate the annotation problem. Nonetheless, unsupervised methods have significantly decreased the accuracy of unsupervised models in video-based person Re-ID compared to supervised methods. In the future, consideration will be given to introducing a new algorithm that can assign suitable pseudo-labels to unlabeled samples or designing a new loss function that facilitates the training of unsupervised models thereby improving the existing unsupervised methods. In addition, effective data augmentation would be effective in improving the overall new performance of all Re-ID methods.

Existing video Re-ID models [135] have made significant progress in recognition abilities, but their successes often rely on large model parameters, making them challenging to deploy effectively on edge devices. Therefore, in the future researchers should focus on developing lightweight video-based person Re-ID models that can overcome the above limitations and enable efficient deployment on resource-constrained edge devices. The advantages of lightweight models in Re-ID are manifold. Firstly, lightweight models have reduced computational and memory requirements, allowing them to run efficiently on edge devices with limited resources. This enables real-time processing and analysis of video streams locally, without relying on a centralized server or cloud infrastructure. Secondly, lightweight models facilitate easy integration with IoT or edge computing frameworks. Their compact size enables seamless deployment on edge devices, enabling intelligent and context-aware video analysis at the edge. This integration unlocks a wide range of applications, including smart surveillance, personalized services, and real-time decision-making. Overall, lightweight models in video Re-ID not only address the challenges of deploying models on edge devices but also pave the way for the convergence of Re-ID with IoT [51] and edge computing, unlocking new possibilities for intelligent and efficient video analysis in various domains.

Although generative models have achieved significant success in the field of video-based person Re-ID, and many effective generative model-based Re-ID methods [98–100] have emerged, these methods only focus on recognition performance, ignore the privacy protection issues. Therefore, in future design of generative models for Re-ID, the incorporation of encryption algorithms or federated learning techniques can address this concern. By introducing encryption algorithms, sensitive information within video data can be protected through secure computation techniques, ensuring privacy

during the model training and inference process. This safeguards the personal identities and attributes of individuals captured in video, aligning with ethical and legal considerations. Additionally, federated learning allows for decentralized model training, where data remains on local devices or edge servers, minimizing the need to share raw data. This approach not only enhances privacy but also addresses concerns regarding data ownership and data security. Integrating encryption algorithms or federated learning techniques into generative models for video Re-ID offers several advantages. Firstly, it ensures that personal information remains confidential, preserving the privacy rights of individuals. Secondly, it enables compliance with privacy regulations and standards, fostering trust and ethical practices. Thirdly, it facilitates the secure deployment of video Re-ID models in sensitive environments, such as public spaces or healthcare facilities, where privacy protection is paramount. By considering privacy concerns and implementing these techniques, future generative models in the field of video Re-ID can strike a balance between recognition performance and privacy preservation, paving the way for responsible and trustworthy video analysis applications.

Ultimately, the Re-ID approach uses metric learning techniques such as Euclidean distance to compute feature similarity, which is time-consuming and slow to retrieve, and is not applicable to practical applications. More research is needed on how to design a new strategy to replace the metric learning strategy. Therefore, further exploration of video-based Re-ID methods remains an interesting area for future research.

6 Conclusion

This paper aims to act as a catalyst for future researchers by reviewing recent advancements in video-based deep person Re-ID methods. Specifically, based on the motivations, existing video-based deep person Re-ID methods are divided into two categories, namely methods for network architecture design and loss function design. Network architecture design-based methods can be further divided into seven different categories, including global feature methods, local feature methods, attention methods, graph methods, hypergraph methods, Transformer methods, and generative methods. Starting from the idea of loss function design, the related works are divided into four categories, including identity loss, verification loss, triplet loss, and OIM loss. In addition, the paper also evaluates the performance of various methods on four popular datasets and suggests some future research directions, such as cross-modal or multi-modal video Re-ID, unsupervised video Re-ID, and integrating video Re-ID with Internet of Things or edge computing applications.

Acknowledgement: I express my sincere gratitude to all individuals who have contributed to this paper. Their dedication and insights have been invaluable in shaping the outcome of this work.

Funding Statement: We acknowledge funding from National Natural Science Foundation of China under Grants Nos. 62101213, 62103165, the Shandong Provincial Natural Science Foundation under Grant Nos. ZR2020QF107, ZR2020MF137, ZR2021QF043.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Jiahe Wang, Xizhan Gao; data collection: Jiahe Wang, Fa Zhu; analysis and interpretation of results: Xizhan Gao, Xingchi Chen; draft manuscript preparation: Jiahe Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials utilized in this review originate from publicly available databases and previously published studies, with proper citations included throughout the text. References to these sources can be found in the bibliography.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.

[2] M. O. Almasawa, L. A. Elrefaei, and K. Moria, "A survey on deep learning-based person re-identification systems," *IEEE Access*, vol. 7, pp. 175228–175247, 2019. doi: 10.1109/ACCESS.2019.2957336.

[3] G. Chen, J. Lu, M. Yang, and J. Zhou, "Learning recurrent 3D attention for video-based person re-identification," *IEEE Trans. on Image Process.*, vol. 29, pp. 6963–6976, 2020. doi: 10.1109/TIP.2020.2995272.

[4] Z. Zhou, Y. Huang, and W. Wang, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2017, pp. 4747–4756.

[5] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE Trans. on Circuits and Syst. For Video Technol.*, vol. 28, no. 10, pp. 2768–2776, 2017.

[6] T. Wang, S. Gong, and X. Zhu, "Person re-identification by video ranking," in *Proc. Comput. Vis.—ECCV 2014: 13th Eur. Conf.*, Zurich, Switzerland, Sep. 6–12, 2014, pp. 688–703.

[7] M. Farenzena, L. Bazzani, and A. Perina, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. 2010 IEEE Comput. Soc. Conf. on Comput. Vis. and Pattern Recognit.*, 2010, pp. 2360–2367.

[8] M. O. Almasawa, L. A. Elrefaei, and K. Moria, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. on Comput. Vis. and Pattern Recognit.*, 2006, pp. 1528–1536.

[9] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Comput. Vis.—ECCV 2008: 10th Eur. Conf. on Comput. Vis.*, Marseille, France, Oct. 12–18, 2008, pp. 262–275.

[10] B. J. Prosser, W. S. Zheng, and S. Gong, "Person re-identification by support vector ranking," *BMVC*, vol. 2, no. 5, pp. 1–11, 2010. doi: 10.5244/C.24.

[11] W. S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 35, no. 3, pp. 653–668, 2012.

[12] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proc. IEEE Int. Conf. on Comput. Vis.*, 2006, pp. 3267–3574.

[13] Y. Wang *et al.*, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 8042–8051.

[14] K. Islam, "Deep learning for video-based person re-identification: A survey," 2023, *arXiv:2023.11332*.

[15] M. Ye, J. Shen, and G. Lin, "Deep learning for person re-identification a survey and outlook," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2021.

[16] D. Cheng, Y. Gong, and S. Zhou, "Person re-identification based on triplet loss and improved hard positive mining in Siamese network," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2016, pp. 1335–1344.

[17] J. You, A. Wu, X. Li, and W. -S. Zheng, "Top-push video-based person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2016, pp. 1345–1353.

[18] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *Proc. IEEE Int. Conf. on Comput. Vis.*, 2015, pp. 3810–3818.

[19] J. Chen, Y. Wang, and Y. Y. Tang, "Person re-identification by exploiting spatio-temporal cues and multi-view metric learning," *IEEE Signal Process. Lett.*, vol. 23, no. 7, pp. 998–1002, Jul. 2016. doi: 10.1109/LSP.2016.2574323.

[20] L. Wu, C. Shen, and A. V. D. Hengel, "Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach," 2016, *arXiv: 1606.01609*.

[21] N. McLaughlin, J. M. Del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2016, pp. 1325–1334.

[22] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. on Comput. Vis.*, 2017, pp. 4733–4742.

[23] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2017, pp. 5790–5799.

[24] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proc. of the AAAI Conf. on Artif. Intell.*, 2018.

[25] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person ReID," 2018, *arXiv: 1805.02104*.

[26] C. -T. Liu, C. -W. Wu, Y. -C. F. Wang, and S. -Y. Chien, "Spatially and temporally efficient non-local attention network for video-based person re-identification," 2019, *arXiv: 1908.01683*.

[27] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2019, pp. 3958–3967.

[28] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan and X. Chen, "VRSTC: Occlusion-free video person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 7183–71922.

[29] Y. Yan *et al.*, "Learning multi-granular hypergraphs for video-based person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2020, pp. 2899–2908.

[30] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen, "Appearance-preserving 3D convolution for video-based person re-identification," in *Proc. Comput. Vis.-ECCV 2020: 16th Eur. Conf.*, Glasgow, UK, Aug. 23–28, 2020, pp. 228–243.

[31] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2020, pp. 10407–10416.

[32] G. Chen, Y. Rao, J. Lu, and J. Zhou, "Temporal coherence or temporal motion: Which is more critical for video-based person re-identification?," in *Proc. Comput. Vis.-ECCV 2020: 16th Eur. Conf.*, Glasgow, UK, Aug. 23–28, 2020, pp. 660–676.

[33] J. Yang, W. -S. Zheng, Q. Yang, Y. -C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2020, pp. 3289–3299.

[34] R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan, "BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2021, pp. 2014–2023.

[35] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, "Watching you: Global-guided reciprocal learning for video-based person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2021, pp. 13334–13343.

[36] A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury and Z. Wu, "Spatio-temporal representation factorization for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2021, pp. 152–162.

[37] J. Liu, Z. -J. Zha, W. Wu, K. Zheng, and Q. Sun, "Spatial-temporal correlation and topology learning for person re-identification in videos," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2021, pp. 4370–4379.

[38] C. Eom, G. Lee, J. Lee, and B. Ham, "Video-based person re-identification with spatial and temporal memory networks," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2021, pp. 12036–12045.

[39] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu and D. Wang, "Pyramid spatial-temporal aggregation for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2021, pp. 12026–12035.

[40] T. He, X. Jin, X. Shen, J. Huang, Z. Chen and X. -S. Hua, "Dense interaction learning for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2021, pp. 1490–1501.

[41] A. Subramaniam, J. Vaidya, M. A. M. Ameen, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention module for video-based computer vision tasks," *Comput. Vis. Image Underst.*, vol. 223, 2022, Art. no. 103532. doi: 10.1016/j.cviu.2022.103532.

[42] J. Wu *et al.*, "CAViT: Contextual alignment vision transformer for video object re-identification," in *Proc. Eur. Conf. on Comput. Vis.*, 2022, pp. 549–566.

[43] F. Yang, X. Wang, X. Zhu, B. Liang, and W. Li, "Relation-based global-partial feature learning network for video-based person re-identification," *Neurocomputing*, vol. 488, pp. 424–435, 2022. doi: 10.1016/j.neucom.2022.03.032.

[44] T. Chai, Z. Chen, A. Li, J. Chen, X. Mei and Y. Wang, "Video person re-identification using attribute-enhanced features," *IEEE Trans. on Circuits and Syst. For Video Technol.*, vol. 32, no. 11, pp. 7951–7966, 2022. doi: 10.1109/TCSVT.2022.3189027.

[45] Y. Yao, X. Jiang, H. Fujita, and Z. Fang, "A sparse graph wavelet convolution neural network for video-based person re-identification," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108708. doi: 10.1016/j.patcog.2022.108708.

[46] X. Liu, C. Yu, P. Zhang, and H. Lu, "Deeply coupled convolution-transformer with spatial-temporal complementary learning for video-based person re-identification," *IEEE Trans. on Neural Netw. and Learn. Syst.*, 2023. doi: 10.1109/TNNLS.2023.3271353.

[47] M. Kim, M. Cho, and S. Lee, "Feature disentanglement learning with switching and aggregation for video-based person re-identification," in *Proc. IEEE/CVF Winter Conf. on Appl. of Comput. Vis.*, 2023, pp. 1603–1612.

[48] H. Tao, Q. Duan, and J. An, "An adaptive interference removal framework for video person re-identification," *IEEE Trans. on Circuits and Syst. For Video Technol.*, 2023. doi: 10.1109/TCSVT.2023.3250464.

[49] J. Leng, H. Wang, X. Gao, Y. Zhang, Y. Wang and M. Mo, "Where to look: Multi-granularity occlusion aware for video person re-identification," *Neurocomputing*, vol. 536, pp. 137–151, 2023. doi: 10.1016/j.neucom.2023.03.003.

[50] J. Peng, G. Jiang, and H. Wang, "Adaptive memorization with group labels for unsupervised person re-identification," *IEEE Trans. on Circuits and Syst. For Video Technol.*, vol. 33, no. 10, pp. 5802–5813, 2023. doi: 10.1109/TCSVT.2023.3258917.

[51] F. Yang, W. Li, B. Liang, and J. Zhang, "Spatiotemporal interaction transformer network for video-based person reidentification in internet of things," *IEEE Internet Things J.*, vol. 10, no. 14, pp. 12537–12547, 2023. doi: 10.1109/JIOT.2023.3250652.

[52] X. Liu, P. Zhang, C. Yu, X. Qian, X. Yang and H. Lu, "A video is worth three views: Trigeminal transformers for video-based person re-identification," *IEEE Trans. on Intell. Transport. Syst.*, 2024. doi: 10.1109/TITS.2024.3386914.

[53] C. Yu, X. Liu, Y. Wang, P. Zhang, and H. Lu, "TF-CLIP: Learning text-free clip for video-based person re-identification," in *Proc. AAAI Conf. on Artif. Intell.*, 2024, pp. 6764–6772.

[54] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognit. Lett.*, vol. 33, no. 14, pp. 1828–1837, 2012. doi: 10.1016/j.patrec.2012.02.014.

[55] R. Satta, "Appearance descriptors for person re-identification: A comprehensive review," 2013, *arXiv: 1307.5748*.
[56] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image Vis. Comput.*, vol. 32, no. 4, pp. 270–286, 2014. doi: 10.1016/j.imavis.2014.02.001.
[57] B. Lavi, M. F. Serj, and I. Ullah, "Survey on deep learning techniques for person re-identification task," 2018, *arXiv: 1807.05284*.
[58] K. Wang, H. Wang, M. Liu, X. Xing, and T. Han, "Survey on person re-identification based on deep learning," *CAAI Trans. on Intell. Technol.*, vol. 3, no. 4, pp. 219–227, 2018. doi: 10.1049/cit2.v3.4.
[59] D. Wu *et al.*, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, 2019. doi: 10.1016/j.neucom.2019.01.079.
[60] H. Masson *et al.*, "A survey of pruning methods for efficient person re-identification across domains," 2019, *arXiv: 1907.02547*.
[61] H. Wang, H. Du, Y. Zhao, and J. Yan, "A comprehensive overview of person re-identification approaches," *IEEE Access*, vol. 8, pp. 270–286, 2014.
[62] X. Lin, P. Ren, Y. Xiao, X. Chang, and A. Hauptmann, "Person search challenges and solutions: A survey," 2021, *arXiv: 2105.01605*.
[63] X. Lin, P. Ren, C. -H. Yeh, L. Yao, A. Song and X. Chang, "Unsupervised person re-identification: A systematic survey of challenges and solutions," 2021, *arXiv: 2109.06057*.
[64] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. on Circuits and Syst. for Video Technol.*, vol. 30, no. 4, pp. 1092–1108, 2019. doi: 10.1109/TCSVT.2019.2898940.
[65] Y. Su, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.
[66] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 7794–7803.
[67] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, 2018, pp. 486–504.
[68] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 6000–6010, 2017.
[69] L. Zhang *et al.*, "Ordered or orderless: A revisit for video based person re-identification," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 43, no. 4, pp. 1460–1466, 2020. doi: 10.1109/TPAMI.2020.2976969.
[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2016, pp. 770–778.
[71] Y. Fu, X. Wang, Y. Wei, and T. Huang, "STA: Spatial-temporal attention for large-scale video-based person re-identification," in *Proc. AAAI Conf. on Artif. Intell.*, 2019, pp. 8287–8294.
[72] Y. Liu, Z. Yuan, W. Zhou, and H. Li, "Spatial and temporal mutual promotion for video-based person re-identification," in *Proc. AAAI Conf. on Artif. Intell.*, 2019, pp. 8786–8793.
[73] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 8786–8793.
[74] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X. -S. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 4913–4922.
[75] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Comput. Vis.-ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Oct. 11–14, 2016, pp. 135–153.
[76] L. Bao, B. Ma, H. Chang, and X. Chen, "Preserving structural relationships for person re-identification," in *Proc. 2019 IEEE Int. Conf. on Multimedia & Expo Workshops (ICMEW)*, 2019, pp. 120–125.
[77] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," in *Proc. Comput. Vis.-ECCV 2020: 16th Eur. Conf.*, Glasgow, UK, Aug. 23–28, 2020, pp. 388–405.

[78] Q. Li, J. Huang, and S. Gong, "Local-global associative frame assemble in video re-ID," 2021, *arXiv: 2110.12018*.
[79] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 1179–1188.
[80] A. Subramaniam, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention networks for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2019, pp. 562–572.
[81] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Trans. on Image Process.*, vol. 228, no. 9, pp. 4192–4205, 2019. doi: 10.1109/TIP.2019.2908062.
[82] M. Kiran, A. Bhuiyan, L. -A. Blais-Morin, M. Javan, I. B. Ayed and E. Granger, "A flow-guided mutual attention network for video-based person re-identification," 2020, *arXiv: 2008.03788*.
[83] P. Hu, J. Liu, and R. Huang, "Concentrated Multi-Grained Multi-Attention Network for Video Based Person Re-Identification," 2020, *arXiv: 2009.13019*.
[84] W. Zhao, Y. Huang, G. Wang, B. Zhang, Y. Gao and Y. Liu, "Multi-scale spatio-temporal feature adaptive aggregation for video-based person re-identification," *Knowl. Based Syst.*, vol. 299, 2024, Art. no. 111980. doi: 10.1016/j.knosys.2024.111980.
[85] S. Bai, B. Ma, H. Chang, R. Huang, and X. Chen, "Salient-to-broad transition for video person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2022, pp. 7339–7348.
[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25. Scotland: Curran Associates, Inc., 2012.
[87] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. on Image Process.*, vol. 28, no. 6, pp. 2976–2990, 2019. doi: 10.1109/TIP.2019.2893066.
[88] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan and X. Yang, "Person re-identification via recurrent feature aggregation," in *Proc. Comput. Vis.-ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Oct. 11–14, 2016, pp. 701–716.
[89] D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann and N. Zheng, "Deep feature learning via structured graph laplacian embedding for person re-identification," *Pattern Recognit.*, vol. 82, pp. 94–104, 2018. doi: 10.1016/j.patcog.2018.05.007.
[90] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 8649–8658.
[91] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu and X. Yang, "Learning context graph for person search," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 2158–2167.
[92] Y. Wu, O. E. F. Bourahla, X. Li, F. Wu, Q. Tian and X. Zhou, "Adaptive graph representation learning for video person re-identification," *IEEE Trans. on Image Process.*, vol. 29, pp. 8821–8830, 2020. doi: 10.1109/TIP.2020.3001693.
[93] D. Chen, A. Doering, S. Zhang, J. Yang, J. Gall and B. Schiele, "Keypoint message passing for video-based person re-identification," in *Proc. AAAI Conf. on Artif. Intell.*, 2022, 239–247.
[94] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, 2018, pp. 399–417.
[95] Z. -M. Chen, X. -S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 5177–5186.
[96] T. Zhang *et al.*, "Spatiotemporal transformer for video-based person re-identification," 2021, *arXiv: 2103.16469*.
[97] C. Zheng, P. Wei, N. Zheng, C. Zheng, P. Wei and N. Zheng, "A duplex spatiotemporal filtering network for video-based person re-identification," in *Proc. 2020 25th Int. Conf. on Pattern Recognit. (ICPR)*, 2021, pp. 7551–7557.

[98] Y. Ge, F. Zhu, D. Chen, R. Zhao, X. Wang and H. Li, "Structured domain adaptation with online relation regularization for unsupervised person re-ID," *IEEE Trans. on Neural Netw. and Learn. Syst.*, vol. 35, no. 1, pp. 258–271, 2022. doi: 10.1109/TNNLS.2022.3173489.

[99] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Trans. on Image Process.*, vol. 29, pp. 1233–1245, 2019. doi: 10.1109/TIP.2019.2940684.

[100] J. Liu *et al.*, "Identity preserving generative adversarial network for cross-domain person re-identification," *IEEE Access*, vol. 7, pp. 114021–114032, 2019. doi: 10.1109/ACCESS.2019.2933910.

[101] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 5177–5186.

[102] M. Ye, X. Zhang, P. C. Yuen, and S. -F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 5177–5186.

[103] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *Proc. 2018 IEEE Winter Conf. on Appl. of Comput. Vis. (WACV)*, 2018, pp. 748–756.

[104] X. Fan, W. Jiang, H. Luo, and M. Fei, "SphereReID: Deep hypersphere manifold embedding for person re-identification," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 51–58, 2019. doi: 10.1016/j.jvcir.2019.01.010.

[105] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *Proc. IEEE/CVF Int. Conf. on Comput. Vis.*, 2019, pp. 4976–4985.

[106] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. on Comput. Vis.*, 2017, pp. 3754–3762.

[107] H. Luo *et al.*, "A strong baseline and batch normneuralization neck for deep person reidentification," 2019, *arXiv: 1906.08332*.

[108] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *Advances in Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019, vol. 32.

[109] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 994–1003.

[110] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2014, pp. 152–159.

[111] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. on Multimedia Comput. Commun. and Appl. (TOMM)*, vol. 14, no. 1, pp. 1–20, 2017.

[112] X. Wang, Y. Hua, E. Kodirov, G. Hu, and N. M. Robertson, "Deep metric learning by online soft mining and class-aware attention," in *Proc. AAAI Conf. on Artif. Intell.*, 2019, pp. 5361–5368.

[113] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv: 1703.07737*.

[114] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, 2018, pp. 402–419.

[115] H. Shi *et al.*, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Comput. Vis.-ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Oct. 11–14, 2016, pp. 732–748.

[116] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2017, pp. 3741–3750.

[117] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, 2018, pp. 188–204.

[118] M. Jiang, B. Leng, G. Song, and Z. Meng, "Weighted triple-sequence loss for video-based person re-identification," *Neurocomputing*, vol. 381, pp. 314–321, 2020. doi: 10.1016/j.neucom.2019.11.088.

[119] S. Kumar, E. Yaghoubi, and H. Proença, "A symbolic temporal pooling method for video-based person re-identification," 2020, *arXiv: 2006.11416*.
[120] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2017, pp. 403–412.
[121] J. Song, Y. Yang, Y. -Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 719–728.
[122] Q. Yang, H. -X. Yu, A. Wu, and W. -S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 3633–3642.
[123] Z. Liu, J. Wang, S. Gong, H. Lu, and D. Tao, "Deep reinforcement active learning for human-in-the-loop person re-identification," in *Proc. IEEE/CVF In. Conf. on Comput. Vis.*, 2019, pp. 6122–6131.
[124] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2017, pp. 3415–3424.
[125] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.*, 2019, pp. 598–607.
[126] P. Pathak, A. E. Eshratifar, and M. Gormish, "Video person re-ID: Fantastic techniques and where to find them (student abstract)," in *Proc. AAAI Conf. on Artif. Intell.*, 2020, pp. 13893–13894.
[127] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Image Anal.: 17th Scand. Conf., SCIA 2011*, Sweden, Springer, May 2011, pp. 91–102.
[128] L. Zheng *et al.*, "Mars: A video benchmark for large-scale person re-identification," in *Proc. Comput. Vis.-ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Springer, Oct. 11–14, 2016, pp. 868–884.
[129] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2018, pp. 1169–1178.
[130] Z. Wang *et al.*, "Robust video-based person re-identification by hierarchical mining," *IEEE Trans. on Circuits and Syst. for Video Technol.*, vol. 32, no. 12, pp. 8179–8191, 2021. doi: 10.1109/TCSVT.2021.3076097.
[131] S. Bai, B. Ma, H. Chang, R. Huang, S. Shan and X. Chen, "SANet: Statistic attention network for video-based person re-identification," *IEEE Trans. on Circuits and Syst. for Video Technol.*, vol. 32, no. 6, pp. 3866–3879, 2021. doi: 10.1109/TCSVT.2021.3119983.
[132] X. Yang, L. Liu, N. Wang, and X. Gao, "A two-stream dynamic pyramid representation model for video-based person re-identification," *IEEE Trans. on Image Process.*, vol. 30, pp. 6266–6276, 2021. doi: 10.1109/TIP.2021.3093759.
[133] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proc. Eur. Conf. on Comput. Vis. (ECCV)*, 2018, pp. 170–186.
[134] M. Wang, B. Lai, Z. Jin, X. Gong, J. Huang and X. Hua, "Deep active learning for video-based person re-identification," 2018, *arXiv: 1812.05785*.
[135] C. -Y. Wang, P. -Y. Chen, M. -C. Chen, J. -W. Hsieh, and H. -Y. M. Liao, "Real-time video-based person re-identification surveillance with light-weight deep convolutional networks," in *Proc. 2019 16th IEEE Int. Conf. on Adv. Video and Signal Based Surveillance (AVSS)*, IEEE, 2019, pp. 1–8.