Video-based person re-identification by intra-frame and inter-frame graph neural network

Guiqing Liu, Jinzhao Wu

PII: S0262-8856(20)30200-6

DOI: https://doi.org/10.1016/j.imavis.2020.104068

Reference: IMAVIS 104068

To appear in: *Image and Vision Computing*

Received date: 11 September 2020

Revised date: 30 October 2020

Accepted date: 1 November 2020

Please cite this article as: G. Liu and J. Wu, Video-based person re-identification by intra-frame and inter-frame graph neural network, *Image and Vision Computing* (2020), https://doi.org/10.1016/j.imavis.2020.104068

Video-based Person Re-identification by Intra-frame and Inter-frame Graph Neural Network

Guiqing Liu[a,b,d], Jinzhao Wu[a,c,d,*]

[a] *Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, Sichuan, 610041, China.*

[b] *College of ASEAN Studies, Guangxi University for Nationalities, Nanning 530006, China.*

[c] *College of Mathematics and Information Science, Guangxi University, Nanning 530004, China.*

[d] *University of Chinese Academy of Sciences, Beijing 100049, China.*

Abstract

In the past few years, video-based person re-identification (Re-ID) have attracted growing research attention. The crucial problem for this task is how to learn robust video feature representation, which can weaken the influence of factors such as occlusion, illumination, and background etc. A great deal of previous works utilize spatio-temporal information to represent pedestrian video, but the correlations between parts of human body are ignored. In order to take advantage of the relationship among different parts, we propose a novel Intra-frame and Inter-frame Graph Neural Network (I2GNN) to solve the video-based person Re-ID task. Specifically, (1) the features from each part are treated as graph nodes from each frame; (2) the intra-frame edges are established by the correlation between different parts; (3) the inter-frame edges are constructed between the same parts across adjacent frames. I2GNN learns video representations by employing the adjacent matrix of the graph and input features to conduct graph convolution, and then adopts projection metric learning on Grassman manifold to measure the similarities between learned pedestrian features. Moreover, this paper proposes a novel occlusion-invariant term to make the part features close to their center, which can relive several uncontrolled complicated factors, such as occlusion and pose invariance. Besides, we have carried

---

[☆] Fully documented templates are available in the elsarticle package on CTAN.

[*] Corresponding author

Email address: wuasean@163.com (Jinzhao Wu)

URL: www.elsevier.com (Guiqing Liu)

out extensive experiments on four widely used datasets: MARS, DukeMTMC-VideoReID, PRID2011, and iLIDS-VID. The experimental results demonstrate that our proposed I2GNN model is more competitive than other state-of-the-art methods.

Keywords: Person re-identification, Graph Neural Network, Intra and Inter Frame, Body Part, Video matching.
2010 MSC: 00-01, 99-00

1. Introduction

Person re-identification (Re-ID) is a very practical research direction, which can be used in monitoring, security, criminal investigation and other fields. Re-ID task means that given a specified person, it retrieves the correct pedestrian in different places, times and cameras. In recent years, Re-ID technology has made great progress, but it has not reached the practical level of face recognition. The main reason is that Re-ID needs to solve problems caused by the resolution of pictures and videos, human posture, scale, occlusion, light and other factors, which make it difficult to capture discriminative features like human face preserved.

The popular solution of person Re-ID is to conduct matching between pedestrian images, including representation learning [1, 2, 3] and distance metric learning [4, 5, 6]. For representation learning, we have transitioned from traditional hand-designed features to deep neural networks, which can exploit more discriminative information. At present, in the task of person Re-ID, people notice more local features in addition to extracting global information of pedestrian images. For example, some hard part methods [7] were adopted in the early stage, and later evolved to adaptive part methods[8], and then to part detection methods. At the mean time, the robustness of the extracted feature also gradually increase. The goal of distance metric learning is to learn an appropriate feature space, where feature vectors from the same person are very close, and the ones belonging to different identities are far apart. For example, [9] proposed a method for simultaneously learning pedestrian features and corresponding pairwise similarity metrics. However, the recognition method based on image can only extract spatial features alone, which cannot solve the occlusion problem. In addition, most of existing person Re-ID models mentioned above obtain features from different pedestrian body parts independently, without considering

their relations among them.

Recently, video-based Re-ID has been taken seriously in the literature, because this technology is more realistic and critical in real-world surveillance applications. In the recent research, the earlier methods were based on image-set, where each frame was treated as an independent image, and the entire video was treated as a group of images, regardless of the relationship between frames. Later, the spatial-temporal feature learning methods are gradually developed, and most methods can be divided into two categories: (i) Optical flow encodes to capture low-order information. Such as, Chung *et al*. [10] proposed a weighted method of two streams structure considering both appearance and optical flow. (ii) Temporal pooling and Recurrent Neural Network (RNN) establish the association between frames to effectively extract high-order information. The methods described above are very effective for solving the video-based person Re-ID task, but there still exists unresolved problems such as occluded frames,pose variations (Figure 1), and these problems lead to degrade the recognition performance of trained models. Those above methods still do not tackle them well. Therefore, how to process frame occlusion is the key problem for video based person Re-ID task. In practical video-based person Re-ID task, if a pedestrian body part from one frame is occluded, we can find the relevant part that is not occluded from other frames. Besides, pedestrian pose in different frames varies a lot, so the pose alignment should be carried out effectively. To achieve Both of above schemes, correlations of pedestrian body parts should be exploited according to natural structure of human body.

Figure 1: Several samples in datasets. This figure illustrates the challenging variations of occlusion, pose, background, illumination, and viewpoints from different camera views in a pedestrian walking cycle. Existing video-based methods reduce the influence of some ambiguous cases like occlusions with more continuous images to learn multiple visual features.

As one of the most excellent networks for structural data representation, graph neural network can update and aggregate the feature of current node according to the information of its neighboring nodes, which has made great achievements in natural language processing, computer vision and other fields. Inspired by the function of graph neural network, we firstly divide the pedestrian image into different parts according to the semantic structure information, and construct

the intra-frame graph, which can be used to pass message across different part features. Then we connect the same parts from adjacent frames to build an inter-frame graph. If occlusion occurs in a frame, the part of the adjacent frame and the correlated part of the self-frame are used for message passing and feature aggregation. Thus, the features aggregated by our method can effectively learn feature representation even there exists occlusion in several video frames.

Based on above observations, a novel Intra-frame and Inter-frame Graph Neural Network (I2GNN) is proposed for video-based person Re-ID task. This method partitions each person as N key parts, (including head, torso, 2 arms and 2 legs), which are extracted local features for them by an intra-frame feature extractor, and the regional features extracted in each part are served as graph nodes. Then, we build the intra-frame graph on local part features for each frame to exploit spatial information. At the same time, I2GNN constructs the inter-frame edges between the same parts among different frames to learn temporal representations. After that, I2GNN learns the final spatio-temporal pedestrian feature representation on the designed intra-frame and inter-frame topology graph by graph neural network. In the end, we adopt projection metric learning to calculate the distance between different pedestrian GNN feature matrices.

1.1. Contributions

The major novelties of this work are summarized as follows:

1. We propose a novel Intra-frame and Inter-frame Graph Neural Network (I2GNN) for video-based person re-identification task. I2GNN constructs an intra-frame and inter-frame messaging graph structure according to the semantic relations of human body parts among intra- and inter- frames. By using the message passed by the graph, current nodes can be better updated and aggregated into graph representations. It can be seen from the experimental results that the method of using adjacent nodes to update current node state is robust. In response to factors such as occlusion in the video frame and changes of human pose, we have designed occlusion-invariant loss to guide our model with relieving negative effects of these factors.

2. We design a uncontrolled factor relive loss and adopt projection metric learning on Grassman Manifold to guide the training of I2GNN model, which is enable to learn discriminative features for each pedestrian video.

3. We have carried out extensive experiments on four Benchmarks (i.e. MARS, PRID2011 iLIDS-VID, and DukeMTMC-VideoReID), and the results elaborate the effectiveness and

superiority of our proposed method, compared to state-of-the-arts.

2. Related work

In this part, we introduce related works of person re-identification(Re-ID), which are mainly divided into two categories of image-based, and video-based person re-identification. Besides, we also review several recent Re-ID models based on Graph Neural Networks (GNN) as the background of this paper.

2.1. Image-based person Re-ID

In recent years, researchers have put forward a large number of models in image-based person re-identification. With the development of convolution neural network, feature extracting methods have been evolved from the earliest manual feature [11, 12, 13] to more robust feature extraction through deep neural network [14, 15, 16]. At present, it is mainly divided into two categories, representation learning [1, 2, 3] and distance metric learning [4, 5, 6].

Representation learning uses various CNN architectures to extract more robust features for identity classification, such as part-based [7, 8, 17] methods. Sun et al [7] proposed a reasonable partition strategy called PCB, which can learn discriminative part-informed features. To solve the problem of misalignment of body parts, Zhao et al [8] proposed a method of part alignment based on attention model. For distance metric learning, Ahmed *et al*. [18] proposed to apply the Siamese structure to deep neural network and share the weights of network parameters, so as to reduce the intra-class gap and increase the inter-class gap. Hermans *et al*. [14] used triplet losses to teach the network pushing features of the same person closer and pulling features of different identity farther. However, the disadvantage of above methods is that the occlusion problem cannot be solved. In our proposed method, we adopt a part-based ideology, which divide human body into 6 key parts (including head, torso, 2 arms and 2 legs), and extract local features as nodes through the intra-frame feature extractor.

2.2. Video-based person Re-ID

Video-based person Re-ID is an extended study of image-based person Re-ID. This section briefly reviews two categories of video person re-identificaition feature learning that are closely

related to this work. (i) In many studies [19, 20], people adopted optical flow to learn temporal features, which encodes to capture low-order information. For example, Simonyan *et al*. [20] introduced a dual-flow network to learn the spatial and temporal features of superimposed optical flow. A potential problem is the sensitivity of optical flow to spatial misadjustment error, and it usually occurs between adjacent human bounding boxes. (ii) Another solution is temporal pooling and Recurrent Neural Network (RNN) to establish the association between frames, which can effectively extract high-order information. Among them, they apply temporal pooling to aggregate feature representations across all timestamps. For example, Li *et al*.[19] utilized partial cues and weights learning strategies to fuse features extracted from video frames. Then, the output of RNN passes through the time pool to extract the video representations. Mclaughlin *et al*. [21] firstly extracted image level features, and used RNN to establish temporal cues as inter-frames model. But many of these methods are still unsuccessful to solve the problem of occluded frames, because they directly learn temporal cues without considering the occluded regions of human body parts. If there exists an occlusion, the feature representation would be influenced by unseen part and the robustness of learned features must be greatly reduced. In this paper, in addition to using the local features of each frame to construct an intra-frame graph, we also establish inter-frame graphs by connecting inter-frame edges between same parts of sequential frames. In a video sequence, not all frames will be blocked, and the information of adjacent nodes is passed through the graph to update the features, so as to reduce the impact of occlusion on the result of robustness of the extracted features.

2.3. GNN-based re-ID Models

The basic principle of graph neural network has excellent capability of supervised training and semi-supervised learning by considering the relationship among samples, which has been gradually applied to some computer vision tasks recently. In recent years, person Re-ID is gradually combined with the GNN models [22, 23, 24, 25, 26]. Cheng et al. [27] proposed a structured graph Laplace embedding algorithm, which makes full use of the structured distance relationship between training samples and expresses the relationship in the form of graph Laplace. PH-GCN [28] combined part-based GCN to extract more compact and robust feature representations. In unsupervised person Re-ID task, Jiang *et al* [28] proposed a new unsupervised graph association (UGA) method, which can reduce the damage of noise association and learn

features from the bottom layer by mining the graph information. STGCN [29] proposes a new network consisting of two GCN branches, in which the spatial branch extracts the structure information of the human body, and the time branch mines discriminant clues from adjacent frames. Through joint optimization of these branches, robust spatio-temporal information complementing the manifestation information can be extracted. Wang *et al*. [30] have viewed the local features of the human body as nodes, and used the Adaptive direction graph convolutional (ADGC) layer to transfer the information between nodes, and introduce the graph matching to be scored according to the predicted similarity, which can achieve the purpose of learning high-order relationship and topology information to distinguish the features and robust alignment.

This part concludes the methods of combining person Re-ID with GCN. In detail, the images or videos are regarded as graph nodes, and the relationship between each node is represented by constructing graph. In our method, we design a new graph neural network to learn the final spatio-temporal pedestrian feature representation. Each part from video frames are regarded as graph nodes, and the graph structure is constructed according to their inter-frame and inter-frame body part relations.

3. Proposed Approach

In this part, we will briefly review the algorithm for Intra-frame and Inter-frame Graph Neural Network (I2GNN). It proposes intra-frame and inter-frame node embedding modules to update the features of each human body part. Besides, the projective metric learning is introduced to calculate the distance between different affinity feature matrices. Furthermore, we integrate an occlusion-invariant loss into I2GNN which aims to overcome the influence of occluded frames.

Figure 2: Overview of our Intra-frame and Inter-frame Graph Neural Network. (1) T frames are sampled from a long-range video with a restricted random sampling method. (2) For each video frame, intra-frame feature extractor is used to extract the Nd-dimension feature, where N represents the number of parts, and the feature vector for each part has d dimensions. (3) The extracted feature vectors are treated as graph nodes, and then employ GNN to perform feature propagation on the graph iteratively in Intra-frame and Inter-rame Embedding Layers. (4) Feature

vectors from Embedding Layers are calculate as affinity metrics. (5) We calculate the distance between two affinity metrics by projection metric learning.

3.1. Network Overview

In the past researches, graph neural network updates the state of the self node by using the information from adjacent nodes. Inspired by those approaches, we propose the I2GNN, which adopt the relationships between intra-frames and inter-frames to solve the problem of uncontrollable factors such as occlusion in pedestrian videos. The network includes intra-frame and inter-frame node embedding, projection metric learning and related loss functions. In this section, the specific content of each module will be given. Firstly, we design a feature extractor to extract the features from human body parts according to the spatial relationships. We utilize the relationship between body parts within and between frames to build the intra-frame graph and the inter-frame graph. Secondly, we embed the graph structure into the intra-frame and inter-frame node embedding to update node features by using the message from adjacent features and self-features. Thirdly, we use affinity matrix to represent the correlations between node and node in a pedestrian video. Then, we adopt projection metric to calculate distance between two affinity feature matrices. At the end, we train our model with contrastive and softmax losses to reduce the distance of intra-identity and increase the distance of inter-identity. This module can effectively calculate the distance between high dimensional features. In order to reduce or eliminate the effect of uncontrolled factors (*e.g.* occluded frames), we develop a video-level occlusion-invariant loss to pulling all features to the overall mean representative characteristics, which has relative robustness to occlusion compared to single frame. Figure 2 is a detailed diagram of our I2GNN method.

3.2. Feature Extraction

In this stage, the goal of our method is to extract the important features of pedestrians and reduce the impact of background clutters, occlusion and varying pose. Several works [8, 7, 28] have verified that the part-based method is effective in person re-identification. The occlusion and other factors in the pedestrian images are noneffective by simple part-based method. Thus, we divide pedestrian images according to the semantic relationship of human body, and construct the graph with human topological structure.Specifically, Following the ideas on person

Re-identification [17], it is effective to divide the human body into six parts according to the semantic relationship and combine with the overall feature.

We take T frames for each person's video sequence in a strictly random sampling manner, and the sequence of video frames is represented as $X = \{x_1, \cdots, x_t, \cdots, x_{T-1}\}$, where T is the length of frames. Then, we utilize pose estimation algorithm [31] to locate key points in the body, which provide cues to partition human body into six parts, as shown in Figure 3.

Figure 3: Part division diagram of human body examples. Each part is in different color box for identification. Since the structure information of human body can provide additional discriminative clues for the transmission of appearance characteristics, the transmission between pedestrian parts can exploit more correlations between video frames. The key point estimation (OpenPose) model is used to extract joint locations from different regions of human body and partition human body parts to learn local feature representations as nodes to formulate graph structure. The graph structure is constructed according to the structure information of human body, and the graph convolutional layer is proposed to transfer the relationship information between nodes.

Then we divide these key points into six parts according to the semantic structure of the human body, including head, torso, right arm, left arms, right leg and left leg. Each part of the human body are defined as $x_t = \{x_t^0, \cdots, x_t^p, \cdots, x_t^P\}$, where x_t^p is the p-th part at the t-th frame x_t in pedestrian video. The feature representation of each part can be expressed by the following mathematical formula:

$$h_t^p = \varphi\left(x_t^p, \theta_b\right) \tag{1}$$

where h_t^p represents feature vector of x_t^p, and $\varphi(\cdot, \theta_b)$ is the backbone convolutional neural network with parameters θ_b. Through Eq. 1, the part feature vectors $\{h_t^0, ..., h_t^p, ..., h_t^P\}$ represent the local characteristics of each part, and h_t represents the global feature of pedestrian image when we take x_t as the input of $\varphi(\cdot, \theta_b)$. We chose ResNet50 pretrained with ImageNet [32] as our backbone CNN feature extractor, which can effectively extract appearance features. Then, pairwise pedestrian input videos are divided as frames and then passed through the network to

obtain their appearance representations.

In order to make the embedding of adjacent nodes together to update the self-embedding of each node, we use the characteristics and affinity of each part to construct adjacency graph. In detail, the node information and edges of the graph are defined as follows:

Node. We define $\mathcal{V} = \{v_1, \cdots, v_i, \cdots, v_{T\times P}\}$ as nodes in our constructed graph for a pedestrian video, and v_i is i-th node in the part sequence. Corresponding to $\mathcal{V}$, the part features are sorted by a uniform sequence of $\mathcal{H} = \{h_1, \cdots, h_i, \cdots, h_{T\times P}\}$ Specifically, the partitioned part from different time-stamps (frames) are divided as head, torso, two arms, and two legs by the constitution of related key joints and represented by $\mathcal{V}$.

Edge. In order to express the relationship among different parts, I2GNN proposes two structures, including inter-frame and intra-frame connected edges. Concretely, the nodes of adjacent body parts in the same frame (Intra-frame) ε_{intra} and the nodes of same body part from adjacent frames (Inter-frame) ε_{inter} are connected which can be formulated by,

$$\varepsilon_{intra}(i,j) = \begin{cases} 1 & i \neq j \text{ and } v_j \in \mathcal{A}_i \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

$$\varepsilon_{inter}(i,k) = \begin{cases} 1 & i \neq k \text{ and } v_j \in \mathcal{T}_i \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

where $\mathcal{A}_i$ denotes the collection of adjacent nodes in same frame for node v_i, and $\mathcal{T}_i$ represents the node set of same part in adjacent frames for node v_i.

3.3. Intra-frame graph node embedding

Based the constructed graph $\{\mathcal{V}, \varepsilon_{inter}\}$ and the feature vectors $\mathcal{H}$ of each part in a pedestrian video, graph neural network can transfer, transform and aggregate the node feature information to update the original spatial region features, and the final graph representations are used for Re-ID task in this paper. Inspired by graph convolutional network [33], we design an inra-frame a graph neural network for the intra-frame graph. For each node v_i, its received message m_{intra}^i and feature vector h_i are updated according to its adjacent nodes at k-th iterations by,

$$m_{intra}^{i}(k)=\frac{1}{\left|(i,j)\in\varepsilon_{intra}\right|}\sum_{j:(i,j)\in\varepsilon_{intra}}\varphi\left(h_j(k-1),\theta_{intra}^{msg}\right) \tag{4}$$

$$v_i(k)=\varphi(h_i(k-1),\theta_{intra}^{node}) \tag{5}$$

$$h_i(k)=\varphi(m_{intra}^{i}(k),v_i(k),\theta_{intra}^{update}) \tag{6}$$

wehre θ_{intra}^{msg}, θ_{intra}^{node}, and θ_{intra}^{update} represent trainable parameters in intra-frame GNN. Eq. (4) defines that the message passing method between adjacent nodes and nodes themselves are passed on the intra-edges. Among them, $\varphi(\cdot,\theta_{intra}^{msg})$ defines the intra-frame message passing function. To avoid unequal number of adjacent nodes for each part, we use the overall adjacent nodes to achieve the normalization of the aggregated features from adjacent nodes. $\varphi\left(\cdot,\theta_{intra}^{node}\right)$ in Eq. (5) defines the self-passing function for each node. Eq. (6) is used to update the current accumulated information according to the messages passed by the adjacent nodes and the self-node. Among the three message passing functions, θ_{intra}^{msg}, θ_{intra}^{node} can be expressed by neural networks with ReLU activation, and θ_{intra}^{update} can be expressed by a summation function.

To make it easily to understand, we summarize the intra-frame graph convolution (IntraGConv) for i=th node between $k-1$ and k layers as,

$$\{\mathcal{H}(k)\}=IntraGConv(\varepsilon_{intra},\mathcal{H}(k-1)) \tag{7}$$

where ε_{intra} is in form of adjacent matrix, and feature nodes $\mathcal{H}$ are embedded into network. Note that, the graph edges ε_{intra} are transformed into the form of $\{0,1\}^{(T\times P)\times(T\times P)}$ as adjacent matrix in Eq. (7).

3.4. Inter-frame node embedding

We have discussed how to update the features in intra-frame graph structure through intra-graph aggregation mentioned above. In this module, we will explore inter-graph aggregation. The formula of this module is similar to that of intra-node embedding, and the main difference is that the message is delivered according to inter-frame edges. Now, we propose the inter-frame node embedding formula as follows,

$$m_{inter}^{i}(k)=\frac{1}{\left|(i,j)\in\varepsilon_{inter}\right|}\sum_{j:(i,j)\in\varepsilon_{inter}}\varphi\left(h_j(k-1),\theta_{inter}^{msg}\right) \tag{8}$$

$$v_i(k) = \varphi(h_i(k-1), \theta_{inter}^{node}) \tag{9}$$

$$h_i(k) = \varphi(m_{inter}^{i}(k), v_i(k), \theta_{inter}^{update}) \tag{10}$$

where θ_{inter}^{msg}, θ_{inter}^{node}, and θ_{inter}^{update} are network parameters in message passing. Besides, Eq. (8) describes that adjacent nodes pass message according to the inter-frame edges, and the normalization of aggregated features is similar to intra-frame GNN module; Eq. (9) represents the self-passing of each node message in the inter-frame graph, and Eq. (10) is used to accumulate massages to update the status of node i. The structures of θ_{inter}^{msg}, θ_{inter}^{node}, and θ_{inter}^{update} keep consistency with that of θ_{intra}^{msg}, θ_{intra}^{node}, and θ_{intra}^{update}. Eq (11) is also adopted to express the inter-fame graph convolution (InterGConv) between layers $k-1$ and k, as summarized by,

$$\{\mathcal{H}(k)\} = InterGConv(\varepsilon_{inter}, \mathcal{H}(k-1)) \tag{11}$$

which denotes a layer of our inter-node embedding net. We also use the form of adjacency matrix ($\varepsilon_{inter} \in \{0,1\}^{(T\times P)\times(T\times P)}$) to represent the inter message passing path.

Based on intra- and inter-frame graph neural networks, the final spatio-temporal feature representation for each pedestrian video can be obtained after last GNN layer, which is calculated by,

$$\{\hat{\mathcal{H}}\} = \mathcal{H}(K) \tag{12}$$

where K is the number of GNN layers, and $\{\hat{\mathcal{H}}\} = \{\hat{h}_1, \cdots, \hat{h}_i, \cdots, \hat{h}_{T\times P}\}$.

3.5. Projection Metric Learning

After graph neural network, our method learns feature vectors for each video frames, and the left problem is how to embedded them into a unified representation to conduct Re-ID. Previous methods directly implement concatenation or computing average for them, that will lose part of identical correlation among them. Another complicated operation is to conduct long-short term memory unit to exploit temporal information, while our GNN has exploited the time-correlations. Thus, both of above methods are inadequacy to handle our GNN feature vectors.

With the help of the above embedding model, the node to node affinity in the embedding space is obtained by encoding the structural affinity between the intra-frame graph and the inter-frame graph. As we all know, the node affinity is not under linear subspaces, which typically lie in a non-Euclidean space of Grassmann manifold. Here, we introduce the projection metric

learning which is an effective technology to measure the similarity metrics on Grassmann manifold. It can simplify the traditional second-order affinity matrix to a linear matrix. Here, we set a bilinear map and an exponential function to make all elements positive in $\hat{\mathcal{H}}$, and obtain its transformation $\mathbf{M}$ for each pedestrian video,

$$\mathbf{M} = \frac{1}{T \times P} \sum_{i=1}^{T \times P} \left(\frac{(\hat{h}_i)^{\mathrm{T}} A \hat{h}_i}{\tau} \right) \tag{13}$$

where the affinity scores between the intra-frame graph and the inter-frame graph are stored in the affinity matrix $\mathbf{M} \in \mathbb{R}^{+N \times N}$ (N is the dimension of feature vector $\hat{h}_i$); A is the learnable weight of affinity function; For $\tau > 0$, with $\tau \to 0^+$, Eq. (13) becomes more discriminative. Moreover, $\mathbf{M}$ is the feature mapping of pedestrian video in higher dimensional space.

To calculate the distance between two mapping of affinity matrices, we adopt Projection Metric Learning on Grassman Manifold to measure their similarity. Employing projection metric learning is inspired by [34], which can achieve geometric perception dimensionality reduction from original Grassman manifolds to lower dimensions and more discriminative Grassmann manifolds. Under the projection mapping $\mathbf{M}\mathbf{M}^T$ framework, the projection matrices $\mathbf{M}\mathbf{M}^T$ is proposed in [35] to represent the elements on the Grassmann manifold. For the pairwise pedestrian videos X_p (Probe) and X_g (Gallery), the inner product is utilized to realize the subspace containing their features, and we calculate their distance metric by,

$$d\left(\mathbf{M}_p\mathbf{M}_p^T, \mathbf{M}_g\mathbf{M}_g^T\right) = 2^{-1/2} \left\| \mathbf{M}_p\mathbf{M}_p^T - \mathbf{M}_g\mathbf{M}_g^T \right\|_F \tag{14}$$

From Eq. (14), where $\mathbf{M}_p$ is affinity matrix of probe video X_p, $\mathbf{M}_g$ is affinity matrix of gallery video X_g, and $\frac{\|\cdot\|}{\|\cdot\|_F}$ denotes the Frobenius normalization.

3.6. LOSS FUNCTIONS

In order to make our model identify specific characters more accurately, we combine three losses to jointly train our network. Firstly, we adopt the contrastive loss function to make the positive distance narrower and the negative distance larger. The key content of contrastive loss layer is to separate feature vectors in pairs and mark the positive and negative of the video pair with 1 or -1. In our training, positive pairs represent two sequences of the same person in different cameras, and negative pairs represent sequences of different people under random cameras.

Specifically, given an input sequence pair (X_p, X_g), their individual feature representations $(\mathbf{M}_p, \mathbf{M}_g)$ are obtained by our I2GNN model. Here, the contrastive loss on I2GNN is calculated by following formula,

$$\mathcal{L}_{con} = \begin{cases} \frac{1}{2} d(\mathbf{M}_p \mathbf{M}_p^T, \mathbf{M}_g \mathbf{M}_g^T), & p = g \\ \frac{1}{2} [\max(\alpha - d(\mathbf{M}_p \mathbf{M}_p^T, \mathbf{M}_g \mathbf{M}_g^T), 0)]^2, & p \neq g \end{cases} \tag{15}$$

where $d\left(\mathbf{M}_p \mathbf{M}_p^T, \mathbf{M}_g \mathbf{M}_g^T\right)$ is the distance metric between X_p and X_g. This objective function encourages features to be close when the pedestrian sequences come from the same person. In contrast, it can make features separated by a margin α when the pairwise videos are from different identities.

Besides, we propose occlusion-invariant loss to address the negative influence from occlusion. It aims to narrow the distance of each part feature to the mean value of same parts from each video frame. The occlusion-invariant function is calculated in following formula,

$$\mathcal{L}_{occ} = \frac{1}{\left|(i,j) \in \varepsilon_{inter}\right|} \sum \left\| h_i - \mu_j \right\| \tag{16}$$

where μ_j represents the mean value of features connected to i-th node in inter-frame graph ε_{inter}. For occluded region, This loss constraint can restore their identical information from the same parts of adjacent frames, so as to eliminate the noisy caused by occlusion and other negative factors.

Finally, the total loss function for each pedestrian video in a batch of our model can be formulated as,

$$\mathcal{L} = \mathcal{L}_{con} + \lambda \mathcal{L}_{occ} \tag{17}$$

where $\lambda \in [0,1]$ controls the importance of $\mathcal{L}_{occ}$.

4. Experiments and Analysis

4.1. Dataset

At present, the dataset of person re-identification is mainly obtained by tracking and monitoring the personnel bounding box in the image or video sequences, which provide sufficient

data for evaluating our I2GNN model. Here, we conduct experiments on two large scale person Re-ID datasets (MARS [36] and DukeMTMC-VideoReID [37]), and two small challenging datasets (iLIDS-VID [38] and PRID2011 [39]).

MARS dataset [36] is the most complete video based person Re-ID dataset at present, including 1261 identity information and nearly 20000 video sequences. These videos are produced through DPM detector and GMMCP tracker, and captured by six cameras, and video sequences from each pedestrian come from at least two camera angles.

DukeMTMC-VideoReID [40] is a subset of the DukeMTMC multi-camera dataset [37], which was collected on outdoor scenario with varying viewpoint, illuminations, background and occlusions using 8 synchronized cameras. It contains 702 identities, each for training and testing, and 408 identities as the distrators. There are 369,656 tracklets for training and 445,764 frames for testing and distractors.

PRID2011 dataset [39] contains personnel videos from two camera views, which contain 385 and 749 identities, respectively. Each personnel video has a different number of frames. We only select persons with a frame number larger than 20.

iLIDS-VID [38] is a small but challenging dataset captured from two non-overlapped camera views, with a total of 600 video sequences from 300 pedestrians. Each person contains two video sequences, which have more than 20 frames.

4.2. Evaluation Metrics

The standard metrics in person Re-ID literature are Cumulative Matching Characteristic (CMC) curves and mean Average Precision(mAP). The CMC curve evaluates the ranking ability of the Re-ID model, and mAP reflects the ranking capability of our model. The setting on MARS and DukeMTMC-VideoReID datasets are follows [41], and we report their rank-n accuracy and mAP results. Besides, we evaluate on PRID2011 and iLIDS-VID data sets according to [38], and divide each dataset into training and test sets, the last accuracy is the average of 100 times cross validation. Since the CMC and mAP of these two data sets are equivalent, only the CMC accuracy is reported in the above two data sets.

4.3. Implementation Details

We achieve the proposed I2GNN method by PyTorch framework, and employ ResNet50 [42] to pre-trained the model on ImageNet as the feature extractor. Then, we set the length and width of all input images to 256×128. In the training stage, we adopt a strict random sampling strategy to extract T = 8 frames to form a sequence. The Adam optimizer is utilized to update the network parameters, with an initial learning rate of 1×10^{-4}, and an initial weight decay of 5×10^{-4}. On each dataset, we trained a total of 300 epochs, and every 100 epoch weight decay decreased by 1/10. Following recommendations in [43], we select 8 identities to form each batch in the network training, and each identity has 4 sequences, thus each batch consists of $8\times4\times8=256$ images. For parameter setting, α in Eq. 15 and λ in Eq.17 are set by 0.45, and 0.4. In the test phase of I2GNN, we use random sampling method to sample T frames of each video, which purpose is to calculate and sort the distance between different video sequences. In order to obtain key points of each image, we use OpenPose [31] as pose estimation algorithm to detect body in video frames. Each pedestrian can be detected into 18 key points, which are divided into 6 parts according to the structural relationship (Figure 3).

4.4. Comparison with State-of-the-art Methods

In order to verify the effectiveness of our proposed I2GNN model, extensive experiments are implemented on MARS, DukeMTMC-VideoReID, PRID2011 and iLIDS-VID, and recent state-of-the-art methods on each dataset are adopted to compare.

Table 1: Re-ID results on MARS dataset. The best results are in **bold** and second ones are in underline.

Models	Rank-1	Rank-5	Rank-20	mAP
BoW+Kissme [36]	30.6	46.2	59.2	15.5
IDE+XQDA [36]	65.3	82.0	89.0	47.6
SpaAtn [44]	82.3	-	-	65.8
Snippet [45]	86.3	94.7	98.2	76.1
STMP [46]	84.4	93.2	96.3	72.7
STA [47]	86.3	95.7	98.1	80.8
GLTR [48]	87.0	95.8	98.2	78.5

COSAM [41]	86.9	95.5	98.0	**87.4**
MG-RAFA [49]	<u>88.8</u>	**97.0**	<u>98.5</u>	85.9
I2GNN(Ours)	**89.1**	<u>96.7</u>	**98.6**	<u>86.3</u>

Performance on MARS. In this section, we compare our approach with several recent methods on MARS dataset. The experimental results of the compared methods on MARS are all from the published works provided by their authors, which is summarized in Table 1. It shows that the accuracy of Rank-1, Rank-5 and Rank-20 obtained by I2GNN model is 89.1%, 96.7% and 98.6% respectively, and the mAP is 86.3%. Table 1 compares I2GNN with several state-of-the-art methods, including BoW+Kissme [36], IDE+XQDA [36], SpaAtn [44], Snippet [45], STMP [46], STA [47], GLTR [48], COSAM [41], and MG-RAFA [49]. The comparison expresses that our I2GNN model achieves the best rank-1 and rank-20 accuracies with increasing of 0.3% and 0.1%, separately, and it obtains second-best results at rank-5 accuracy mAP with small gaps of 0.3% and 1.1%. From the results in Table 1, our I2GNN reaches the best performance in overall comparison to the state-of-the-art methods, which elaborates the considerable contributions our Intra-frame and Inter-frame GNN module and occlusion invariant loss.

Table 2: Re-ID results on DukeMTMC-VideoReID dataset. The best results are in **bold** and second ones are in <u>underline</u>.

Models	Rank-1	Rank-5	Rank-20	mAP
ETAP-Net [40]	83.6	94.6	97.6	78.3
STA [47]	96.2	99.3	-	94.9
COSAM [41]	95.4	99.3	99.8	94.1
GLTR [48]	96.3	99.3	99.7	93.7
I2GNN (Ours)	96.5	99.3	99.7	94.9

Performance on DukeMTMC-VideoReID. To further evaluate the superiority of I2GNN, we also conduct experiments on DukeMTMC-VideoReID dataset [37], compared with four recently proposed methods, including ETAP-Net[40], STA [47],COSAM [41], and GLTR [50]. The results on DukeMTMC-VideoReID dataset are reported in Table 2, and it reveals that I2GNN performs Rank-1 accuracy of 96.5% and mAP of 94.9. The comparison between I2GNN

and other methods explains that I2GNN is better than others and its superiority is illustrated by every evaluated metrics. Thus, the proposed modules of intra-frame and inter-frame GNN and projection metric learning in this paper contribute development on video-based person re-identification.

Table 3: Re-ID results on small datasets of PRID2011 and iLIDS-VID. The best results are in **bold** and second ones are in underline.

Dataset	PRID2011		iLIDS-VID	
Models	Rank-1	Rank-5	Rank-1	Rank-5
SpaAtn [44]	93.2	-	80.2	65.8
Snippet [45]	93.0	99.3	85.4	96.7
STMP [46]	92.7	98.8	84.3	96.8
GLTR [48]	95.5	**100**	86.0	**98.0**
COSAM [41]	-	-	79.6	95.3
MG-RAFA [49]	95.9	99.7	88.6	**98.0**
I2GNN	**96.0**	**100**	**88.9**	96.9

Performance on PRID2011 and iLIDS-VID. To elaborate that I2GNN not only works well on large scale datasets (MARS and DukeMTMC-VideoReID) but also is effective on small datasets, we adopts PRID2011 and iLIDS-VID to conduct experiments, which results are reported in Table 3. The proposed I2GNN model produces rank-1 accuracy of 96.0% and 88.9% on PRID2011 and iLIDS-VID datasets, respectively, and it is obviously more excellent to other models, compared with the state-of-the-art methods. It is shown that, our method presents competitive performance on rank-1 and rank-5 accuracy. From the comparison on these two small datasets, we can summarize that, especially among video person Re-ID datasets, the video representations learned by our method are more representative through the intra-frame and inter-frame node embedding approach.

Through the comparison on both of small and larges scale datasets, we can conclude that, the distinguishing features of intra-frame and inter-frame node embedding can be learned by our method. I2GNN method can exploit intra-frame and inter frame correlations to neutralize the negative influence from occlusion or other factors, which is perfectly solve by our proposed

occlusion-invariant loss constraint.

4.5. Ablation Study

In this paper, we propose a novel Intra-frame and Inter-frame Graph Neural Network (I2GNN) for video-based person Re-ID task, which utilizes occlusion-invariant loss to avoid the influence of negative factors. In order to verify the effectiveness of each component, we modify I2GNN into several methods to demonstrate the improvements caused by the validated components and employ the results on MARS dataset to illustrate their performance, as shown in Table 4.

Figure 4: Rank-1 accuracy and mAP performance of I2GNN and modified methods on MARS dataset.

Analysis of intra- and inter- frame node embedding module

In Section III, we propose an intra-frame and inter-frame node embedding modules to update the original region features iteratively from adjacent nodes. In order to evaluate the effectiveness of intra-frame and the inter-frame node embedding modules, we set up two groups of comparative experiments. A modified experiment sets all the inner edges of intra frame graph to 0, that is, the intra frame node embedding module is removed, which only utilize inter-frame correlations and is defined as Inter-frame Graph Neural Network (IGNN). Another modified experiment is to remove inter-frame node embedding module, by setting all inter-frame edges to 0, which only introduce intra-frame correlations and is named by Only Intra-frame Graph Neural Network (OIGNN).

The results of IGNN and OIGNN can be seen in Figure. 4. Compared to the performance of I2GNN, the Rank-1 accuracy of IGNN on MARS dataset is 74.91%, mAP is 71.20%, and OIGNN achieves rank-1 accuracy of 69.54%, and mAP of 58.92%. The experimental results show that it is effectiveness to update and aggregate features using inter-frame nodes and intra-frame nodes.

Table 4: Results of modified methods on MARS dataset (rank-1 accuracy and mAP).

Models	Rank-1	rank-5	Rank-20	mAP

baseline	83.9	94.2	96.6	76.7
EDI2GNN	80.8	88.2	93.5	72.9
I2GNN	89.1	96.7	98.6	86.3

Analysis of distance metric module

As for validating the Projection Metric Learning module, we replace it by average Euclidean distance among intra- and inter-frame GNN features, which is calculated by,

$$d = \frac{1}{T \times P} \sum_{i=1}^{T \times P} \| M_p(i) - M_g(j) \|^2 \tag{18}$$

where the M_p and M_g represent the feature matrix from probe and gallery videos, respectively. This modified method is called Euclidean Distance Intra-frame and Inter-frame Graph Neural Network (EDI2GNN). Figure 4 also shows the performance of this modified method and compares it to original I2GNN on MARS dataset.

From the comparison of EDI2GNN, we can find that the rank-1 accuracy of EDI2GNN is 80.84%, and the mAP is 72.96%. The rank-1 accuracy of I2GNN is higher at least 8.26%. It is concluded that the Projection Metric Learning on Grassmann Manifold is feasible for our extracted Intra- and Inter- frame GNN feature vectors in video-based person Re-ID task.

Analysis on the effectiveness of $\mathcal{L}_{occ}$

Under the condition of partial occlusion, the personal information contained in the video will be lost, which will decrease the accuracy of video Re-ID. To solving this problem, we propose occlusion-invariant loss $\mathcal{L}_{occ}$ to drive our model with learning feature representations regardless of occluded pedestrian regions. We defined the model as baseline when we train the network only by contrastive loss $\mathcal{L}_{con}$. In order to verify the effect of L_{occ}, we implement baseline on MARS dataset without $\mathcal{L}_{occ}$ to train I2GNN, which results are reported in Table 4.

It can be observed from Table 4 that baseline generates 83.9% rank-1 accuracy and 76.7 mAP on MARS dataset, and occlusion-invariant loss improves the performance of I2GNN at least 5.2% rank-1 accuracy, and 9.6% mAP. The main reason is that $\mathcal{L}_{occ}$ can provide effective learning guidance to mitigate the impact of occlusion on exploiting spatio-temporal feature representation among video frames.

Evaluation of different part divisions.

The human body division strategy plays an important role in I2GNN approach, and this method achieves the best performance when we partition human body by six parts, including head, torso, 2 arms and 2 legs. To evaluate the influence from different number of part divisions, we implement another two partition strategies of 2 parts (top and bottom of the person) and 4 parts (top-left, top-right, bottom-left and bottom-right) on MARS and DukeMTMC-VideoReID datasets. The rank-1 accuracy and mAP results with different number of parts are summarized in Table 5. It is observed that the division of six parts (head, torso, 2 arms and 2 legs) produces the largest rank-1 accuracy and mAP. The division of 2 parts achieves the worst performance, compared to other settings. This evaluation states that the growing number of body parts can raise the performance of our I2GNN model, and the division of 6 parts can learn more locally semantic representations from pedestrian videos, compared to partitioning body into fewer parts.

Table 5: Evaluation for different number of parts on DukeMTMC-VideoReID and MARS datasets.

Dataset	MARS		DukeMTMC	
Part Number	Rank-1	mAP	Rank-1	mAP
2	87.0	84.9	94.2	93.5
4	88.8	85.1	95.1	94.3
6	89.1	86.3	95.6	94.9

Evaluation of different edge weights in GNN

In our Intra-frame and Inter-frame graph Neural Network, the definition of edge weights for constructing adjacent matrices is a major process to operate GNN on frame-level features. The natural human body structure is to construct edge weights in this paper, and here we introduce alternative definitions of confusion matrix to achieve I2GNN, including K Nearest Neighbors (KNN), ε-Radius, and Fully Correlated (FC) graph constructing rules. Their results are summarized in Table 6.

From the comparison of these definitions of edge weights, the graph following temporal body part cues in our method performs the best performance among them, and KNN achieves the

worst results because it is too simple to construct adjacent matrix and the more complicated definition of ε-Radius obtains better Re-ID consequence. By results from KNN, ε-Radius and our temporal body part cues, it can be seen that more complicated graph building algorithm will generate more accurate matching results. However, the fully correlated graph building method achieves lower performance compared to ε-Radius algorithm, because it is so intricate that increases computation complexity in GNN. Thus, the edge weights following temporal body part provide reasonable adjacent matrix in our intra-frame and inter-frame graph neural network and it performs significant Re-ID effectiveness on video-based person Re-ID task.

Table 6: Alternative experiments for different definition of edge weights.

Dataset	MARS		DukeMTMC	
Edge Weight	Rank-1	mAP	Rank-1	mAP
KNN	84.6	82.2	92.4	91.5
ε-Radius	87.9	84.6	94.2	93.3
FC	85.6	83.5	93.8	92.6
Ours	89.1	86.3	95.6	94.9

5. conclusion

This paper proposes a novel Intra-frame and Inter-frame Graph Neural Network (I2GNN) for video-based person re-recognition (Re-ID), which employs the massage passed by the adjacent features among intra-frame and inter-frame correlations by GNN. This method can construct an adjacency graph based on the relationship between each part with correlations in intra and inter frames. Each part is aggregated by the information passed by the adjacency graph, and the intrinsic affinity structure information between the aggregated feature nodes is adaptively obtained. Finally, the affinity matrix is passed to the Grassmann manifold for optimization. In order to mitigate the impact of invariable factors such as occlusion on performance, we propose the occlusion-invariant loss to restore the information of occluded regions from its adjacent frames. We conduct extensive evaluations on four video-based person re-identification datasets, and experimental results confirm the effectiveness of our I2GNN. Furthermore, a series of modified experiments verify the feasibility of each components in our network.

Acknowledgements

Acknowledgements This research has been financed by the National Natural Science Foundation of China Error analysis and control of semi-algebraic model detection method (61772006), the Science and Technology Major Project of Guangxi Research and Application Demonstration of Key Technologies for Intelligent Ship Networking in Beibu Gulf (AA17204096), the Key Research and Development Project of Guangxi DPA-proof full asynchronous RSA security crypto chip: design methods, tools and prototypes (AB17129012), the SpecialFundforBaguiScholars of GuangxiControl system design and verification (2017), and the Promotion Project of BasicFaculties forYoung and Middle aged College Teachers in Guangxi Common Sense Dynamic Logic Reasoning and Its Application (2018KY0164).

References

[1] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1622–1634. doi:10.1109/TPAMI.2012.246.
URL https://doi.org/10.1109/TPAMI.2012.246

[2] C. Liu, S. Gong, C. C. Loy, X. Lin, Person re-identification: What features are important?, in: Computer Vision - ECCV 2012. Workshops and Demonstrations - Florence, Italy, October 7-13, 2012, Proceedings, Part I, 2012, pp. 391–401. doi:10.1007/978-3-642-33863-2_39.
URL https://doi.org/10.1007/978-3-642-33863-2_39

[3] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, IEEE Computer Society, 2014, pp. 144–151. doi:10.1109/CVPR.2014.26.
URL https://doi.org/10.1109/CVPR.2014.26

[4] S. Liao, S. Z. Li, Efficient PSD constrained asymmetric metric learning for person re-identification, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp.

3685–3693. doi:10.1109/ICCV.2015.420.
URL https://doi.org/10.1109/ICCV.2015.420

[5] F. Xiong, M. Gou, O. I. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: D. J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII, Vol. 8695 of Lecture Notes in Computer Science, Springer, 2014, pp. 1–16. doi:10.1007/978-3-319-10584-0_1.
URL https://doi.org/10.1007/978-3-319-10584-0_1

[6] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, CoRR abs/1503.01543. arXiv:1503.01543.
URL http://arxiv.org/abs/1503.01543

[7] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline), in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV, 2018, pp. 501–518. doi:10.1007/978-3-030-01225-0_30.
URL https://doi.org/10.1007/978-3-030-01225-0_30

[8] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 3239–3248. doi:10.1109/ICCV.2017.349.
URL https://doi.org/10.1109/ICCV.2017.349

[9] E. Ahmed, M. J. Jones, T. K. Marks, An improved deep learning architecture for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 3908–3916. doi:10.1109/CVPR.2015.7299016.
URL https://doi.org/10.1109/CVPR.2015.7299016

[10] D. Chung, K. Tahboub, E. J. Delp, A two stream siamese convolutional neural network for person re-identification, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 1992–2000. doi:10.1109/ICCV.2017.218.
URL https://doi.org/10.1109/ICCV.2017.218

[11] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification, in: British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings, 2011, pp. 1–11. doi:10.5244/C.25.68.
URL https://doi.org/10.5244/C.25.68

[12] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 2197–2206. doi:10.1109/CVPR.2015.7298832.
URL https://doi.org/10.1109/CVPR.2015.7298832

[13] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, pp. 2360–2367. doi:10.1109/CVPR.2010.5539926.
URL https://doi.org/10.1109/CVPR.2010.5539926

[14] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, CoRR abs/1703.07737. arXiv:1703.07737.
URL http://arxiv.org/abs/1703.07737

[15] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 3820–3828. doi:10.1109/ICCV.2017.410.
URL https://doi.org/10.1109/ICCV.2017.410

[16] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 1249–1258. doi:10.1109/CVPR.2016.140.
URL https://doi.org/10.1109/CVPR.2016.140

[17] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 907–915.

doi:10.1109/CVPR.2017.103.

URL https://doi.org/10.1109/CVPR.2017.103

[18] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

URL https://doi.org/10.1109/CVPR.2006.100

[19] S. Li, S. Bak, P. Carr, X. Wang, Diversity regularized spatiotemporal attention for video-based person re-identification, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 369–378. doi:10.1109/CVPR.2018.00046.

URL

http://openaccess.thecvf.com/content_cvpr_2018/html/Li_Diversity_Regularized_Spatiotemporal_CVPR_2018_paper.html

[20] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 568–576.

URL

http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos

[21] N. McLaughlin, J. M. del Rincón, P. C. Miller, Recurrent convolutional network for video-based person re-identification, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 1325–1334. doi:10.1109/CVPR.2016.148.

URL https://doi.org/10.1109/CVPR.2016.148

[22] A. Barman, S. K. Shah, Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 1124–1133. doi:10.1109/ICCV.2017.127.

URL https://doi.org/10.1109/ICCV.2017.127

[23] Y. Shen, H. Li, S. Yi, D. Chen, X. Wang, Person re-identification with deep similarity-guided graph neural network, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV, 2018, pp. 508–526. doi:10.1007/978-3-030-01267-0_30.
URL https://doi.org/10.1007/978-3-030-01267-0_30

[24] D. Chen, D. Xu, H. Li, N. Sebe, X. Wang, Group consistent similarity learning via deep CRF for person re-identification, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 8649–8658. doi:10.1109/CVPR.2018.00902.
URL
http://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Group_Consistent_Similarity_CVPR_2018_paper.html

[25] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, X. Yang, Learning context graph for person search, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2019, pp. 2158–2167. doi:10.1109/CVPR.2019.00226.
URL
http://openaccess.thecvf.com/content_CVPR_2019/html/Yan_Learning_Context_Graph_for_Person_Search_CVPR_2019_paper.html

[26] M. Ye, A. J. Ma, L. Zheng, J. Li, P. C. Yuen, Dynamic label graph matching for unsupervised video re-identification, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 5152–5160. doi:10.1109/ICCV.2017.550.
URL https://doi.org/10.1109/ICCV.2017.550

[27] D. Cheng, Y. Gong, X. Chang, W. Shi, A. G. Hauptmann, N. Zheng, Deep feature learning via structured graph laplacian embedding for person re-identification, Pattern Recognition 82 (2018) 94–104. doi:10.1016/j.patcog.2018.05.007.
URL https://doi.org/10.1016/j.patcog.2018.05.007

[28] B. Jiang, X. Wang, B. Luo, PH-GCN: person re-identification with part-based hierarchical graph convolutional network, CoRR abs/1907.08822. arXiv:1907.08822.
URL http://arxiv.org/abs/1907.08822

[29] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, Q. Tian, Spatial-temporal graph convolutional network for video-based person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3289–3299.

[30] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, High-order information matters: Learning relation and topology for occluded person re-identification, CoRR abs/2003.08177. arXiv:2003.08177.
URL https://arxiv.org/abs/2003.08177

[31] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.

[32] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, 2009, pp. 248–255. doi:10.1109/CVPRW.2009.5206848.
URL https://doi.org/10.1109/CVPRW.2009.5206848

[33] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017.
URL https://openreview.net/forum?id=SJU4ayYgl

[34] Z. Huang, R. Wang, S. Shan, X. Chen, Projection metric learning on grassmann manifold with application to video based face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 140–149. doi:10.1109/CVPR.2015.7298609.
URL https://doi.org/10.1109/CVPR.2015.7298609

[35] A. Edelman, T. A. Arias, S. T. Smith, The geometry of algorithms with orthogonality constraints, SIAM J. Matrix Analysis Applications 20 (2) (1998) 303–353. doi:10.1137/S0895479895290954.
URL https://doi.org/10.1137/S0895479895290954

[36] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, MARS: A video benchmark for large-scale person re-identification, in: Computer Vision - ECCV 2016 - 14th European

Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI, 2016, pp. 868–884. doi:10.1007/978-3-319-46466-4_52.
URL https://doi.org/10.1007/978-3-319-46466-4_52

[37] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision, Springer, 2016, pp. 17–35.

[38] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV, 2014, pp. 688–703. doi:10.1007/978-3-319-10593-2_45.
URL https://doi.org/10.1007/978-3-319-10593-2_45

[39] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Image Analysis - 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings, 2011, pp. 91–102. doi:10.1007/978-3-642-21227-7_9.
URL https://doi.org/10.1007/978-3-642-21227-7_9

[40] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5177–5186.

[41] A. Subramaniam, A. Nambiar, A. Mittal, Co-segmentation inspired attention networks for video-based person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 562–572.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
URL https://doi.org/10.1109/CVPR.2016.90

[43] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, L. Lin, SCAN: self-and-collaborative attention network for video person re-identification, IEEE Trans. Image Processing 28 (10) (2019) 4870–4882. doi:10.1109/TIP.2019.2911488.
URL https://doi.org/10.1109/TIP.2019.2911488

[44] S. Li, S. Bak, P. Carr, X. Wang, Diversity regularized spatiotemporal attention for

video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 369–378.

[45] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 1169–1178. doi:10.1109/CVPR.2018.00128.
URL http://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Video_Person_Re-Identification_CVPR_2018_paper.html

[46] Y. Liu, Z. Yuan, W. Zhou, H. Li, Spatial and temporal mutual promotion for video-based person re-identification, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, 2019, pp. 8786–8793. doi:10.1609/aaai.v33i01.33018786.
URL https://doi.org/10.1609/aaai.v33i01.33018786

[47] Y. Fu, X. Wang, Y. Wei, T. Huang, Sta: Spatial-temporal attention for large-scale video-based person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8287–8294.

[48] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, CoRR abs/1908.10049. arXiv:1908.10049.
URL http://arxiv.org/abs/1908.10049

[49] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, CoRR abs/2003.12224. arXiv:2003.12224.
URL https://arxiv.org/abs/2003.12224

[50] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3958–3967.