Spatial-Temporal Graph Convolutional Network for Video-based Person Re-identification

Jinrui Yang[1,3] , Wei-Shi Zheng[1,2,3*] , Qize Yang[1,3] , Yingcong Chen[4] , and Qi Tian[5]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen 518005, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4] The Chinese University of Hong Kong, China
[5]The Huawei Noah's Ark Lab, China
{yangjr27,yangqz}@mail2.sysu.edu.cn,wszheng@ieee.org,yingcong.ian.chen@gmail.com
,tian.qi1@huawei.com

Abstract

While video-based person re-identification (Re-ID) has drawn increasing attention and made great progress in recent years, it is still very challenging to effectively overcome the occlusion problem and the visual ambiguity problem for visually similar negative samples. On the other hand, we observe that different frames of a video can provide complementary information for each other, and the structural information of pedestrians can provide extra discriminative cues for appearance features. Thus, modeling the temporal relations of different frames and the spatial relations within a frame has the potential for solving the above problems. In this work, we propose a novel Spatial-Temporal Graph Convolutional Network (STGCN) to solve these problems. The STGCN includes two GCN branches, a spatial one and a temporal one. The spatial branch extracts structural information of a human body. The temporal branch mines discriminative cues from adjacent frames. By jointly optimizing these branches, our model extracts robust spatial-temporal information that is complementary with appearance information. As shown in the experiments, our model achieves state-of-the-art results on MARS and DukeMTMC-VideoReID datasets.

1. Introduction

The widely deployed close-circuit television cameras produce a mass of visual surveillance data everyday. This brings the necessities of automatic surveillance video understanding. Because of the privacy or economic issue, there is always non-overlapping regions. Thus it is challenging to conduct tracking or activity analysis on a non-overlapping camera network. In this regard, it is crucial to re-identify pedestrians across non-overlapping camera views, which is called person re-identification (Re-ID).

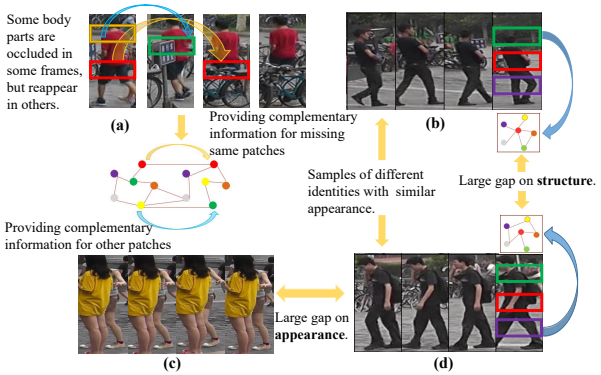

Figure 1. Four image sequences from **four different pedestrians** on MARS dataset. In (a), it is easy to find the same body parts are occluded in some frames but reappear in others. It's easy to differentiate the pedestrians of (c) and (d) due to their large gap in appearance. However, only using the appearance features are not discriminative enough to distinguish (b) and (d), but their structural information of the body is different. Exploiting spatial-temporal relations of parts among the sequence can alleviate these problems.

Existing Re-ID methods can be divided into two categories, i.e., image-based [10, 13, 31, 29, 36, 74, 60, 27, 1, 56, 59, 51, 30, 68, 69, 49] and video-based [8, 4, 61, 11]. Image-based Re-ID takes one or several images as input without considering the temporal information. Generally, it heavily relies on appearance features that are related to color/texture of clothes. When the bounding boxes are not perfect, or there exist noise or occlusions, appearance-based features could be less effective, and image-based Re-ID may not work well in this case. In contrast, by taking a

*Corresponding author

short video clip as input, video-based Re-ID can leverage much richer information, which is potentially beneficial to alleviate the limitations of appearance-based features. To this end, most of the video-based methods employ 3D-CNN [24, 42, 50], RNN/LSTM [63, 66, 75], or attention mechanism [11, 75, 61, 33, 26, 23] to exploit temporal relations from the video. However, these methods only model the temporal relations across the different frames and ignore the potential relations of different parts of the body within a frame or across frames, which may contain more discriminative and robust information for person Re-ID.

As shown in Figure 1 (a), different body parts of persons are occluded or misaligned in different frames, which often causes a performance degeneration for person Re-ID. However, we can also observe some patches of the pedestrian are occluded in some frames but reappear in others. These patches could mutually provide complementary information for each other if we explicitly exploit the temporal relations of patches across different frames, so as to alleviate the occlusion and misalignment problem.

On the other hand, it is effortless to distinguish the pedestrians in Figure 1 (c) and (d) because their appearances are significantly dissimilar. However, the pedestrians in Figure 1 (b) and (d) are very visually similar and the appearance may be not powerful enough for distinguishing in this case. But their structural information of body (e.g., the shape of body) is obviously different which can be seen as complementary with the appearance features and benefits the identification. Thus, capturing the structural information by modeling the spatial relations of patches for each frame is also important.

Inspired by the strong ability of automatically relation modeling of graph convolutional network (GCN) [20] and the successful application of GCN in computer vision, we propose to use GCN to model the relations of different patches. Specifically, we construct the graph by connecting all patches of the different frames to model the temporal relations, aiming at providing complementary information across different patches which can alleviate the occlusion and misalignment problem. On the other hand, we also consider the structural information of intra-frame to provide complementary information of appearance by constructing a patch graph for each frame in a video. Finally, a unified framework, namely Spatial-Temporal Graph Convolutional Network (STGCN), is proposed to simultaneously model the spatial and temporal relations of patches in a video. While graph modeling has been seen in person Re-ID, these methods build a graph on image-level [64, 45] or ignore the structural information within an image [57].

In summary, our contributions are the following. (1) We employ GCN to model the potential relations of different parts of the body within a frame and across frames, providing more discriminative and robust information for person Re-ID. (2) We propose a unified framework that jointly considers the temporal and structural relations and is able to perform end-to-end training. Extensive experiments show that our proposed method outperforms existing state-of-the-art methods on two large-scale video-based person Re-ID datasets.

2. Related Work

Image-based person Re-ID. Existing image-based person Re-ID mainly focus on designing discriminative handcrafted feature [10, 13, 31, 29, 36, 74, 60], distance metric learning [13, 29, 60, 55, 41, 21, 72, 38, 40, 28, 35, 39, 6, 73, 65, 67, 54, 3] or deep learning [27, 1, 56, 59, 51, 30, 68, 69, 49]. However, a video contains many frames and the temporal information is important. Image-based methods lack modeling the temporal relations of a video so these methods are sub-optimal for video-based person Re-ID.

Video-based person Re-ID. Most of video-based methods use optical flow [8, 4, 37, 61], recurrent neural networks (RNNs), temporal pooling [71], or spatial-temporal attention to model the temporal information. Specifically, In [8, 4, 37, 61], the authors use optical flow by computing between adjacent frames to extract temporal features for person Re-ID. However, the process of computing optical flow is time-consuming and the optical flow is not robust enough for occlusion and noisy. As for RNN-based methods [8, 37, 66, 4, 61, 75], as mentioned in [12], RNN has a limited effect on modeling temporal information in Re-ID task, or too hard to train caused by its complicated structure. Compared with temporal pooling [71] which assigns the same weights to all frames, many attention-based methods [11, 75, 61, 33, 26] learn the weight of different frames or parts from a static perspective, i.e. considering the spatial attention and temporal attention separately. Thus, these methods do not fully consider the temporal relations of body parts across different frames and the effect is limited.

Graph neural network methods. In recent years, graph convolutional networks (GCNs) and its variants [20, 5, 14, 18] have been successfully applied to some tasks in computer vision, such as skeleton-based action recognition [62], video classication [52], and multi-label image recognition [7] due to its strong ability of relations modeling. Similarly, many works [57, 64, 45] also apply GCNs on person Re-ID. Specifically, Yan *et al*. [64] and Shen *et al*. [45] build the graph model on image-level, i.e. each node of the graph represents an image, to consider the relations of among images. However, these methods are image-based which do not consider the temporal relations. Furthermore, they ignore the relations of different body parts of intra-frame or inter-frame. Particularly, Wu *et al*. [57] introduce a graph neural network to enable the contextual interactions between the relevant regional features by exploiting pose alignment connection and feature affinity connection. How-

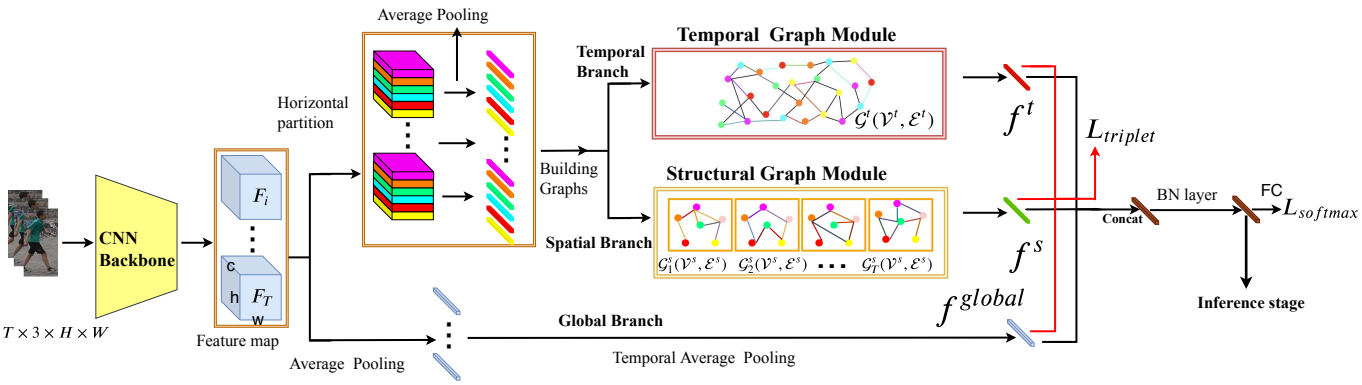

Figure 2. The overall architecture of our proposed method. The input video has T frames and we use a CNN backbone to extract the feature map for each frame. Then, the model is divided into three branches. For the temporal branch and spatial branch, we divide the feature maps horizontally into P patches. These patches are served as nodes in the graph. The temporal branch consists of a temporal GCN module that constructs a graph for each video to model the temporal relations of different patches across different frames. The spatial branch consists of a structural GCN module that constructs graphs for each frame in the video to model the spatial relations of patches within a frame. In the global branch, we perform average pooling on each feature map and then use a temporal average pooling to aggregate features across different frames.

ever, the pose information extraction requires extra computation and it is not integrated into the whole network to perform end-to-end training, which may cause sub-optimal result. Furthermore, this method connects features of different parts of all frames and does not model the spatial relations of body parts for each frame, ignoring the intra-frame structural information.

Compared to these methods, we propose a unified spatial-temporal GCN framework to jointly model the relations of the whole patches in video level and the structural information of individual frames in frame level, which can learn the discriminative and robust spatial-temporal relations of patches to facilitate video-based Re-ID.

3. The Proposed Method

As shown in Figure 2, the architecture for our proposed model consists of three branches. The upper branch is the temporal branch for extracting temporal cues from the patches across the adjacent frames. The middle branch is the spatial branch for extracting structural information of a human body by modeling the spatial relations of patches. The bottom branch is global branch for extracting appearance feature of the pedestrian.

In the following sections, we firstly introduce the construction of patch graph in Section 3.2. Based on this, we further develop the temporal GCN module in Section 3.3 and the structural GCN module in Section 3.4.

3.1. Preliminary

Given a video, we denote it as $V = \{I_1, I_2, \ldots, I_T\}$, where T is the number of frames sampled from the video. For each frame of the video, we denote feature map of frames extracted by the backbone model as F,

$$F = \{F_1, F_2, \ldots, F_T\}, \tag{1}$$

where F_i is the feature map of i-th frame in the video, $F_i \in \mathbb{R}^{h \times w \times c}$, in which h, w, c denotes the height, width and channel number, respectively. Each feature map F_i is horizontally partitioned into P patches. Then each patch is processed by average pooling and each patch feature is represented as x_i. Thus, for a video with T frames, the total number of patches is $N = T \cdot P$. We denote the patches of the video as $p_i = 1, \ldots, N$ and the obtaining patch feature vector as $x_i \in \mathbb{R}^c, i = 1, \ldots, N$.

3.2. Patch Graph Construction

To explore and utilize the relations of patches, we use GCN to model the relations between patches. Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be the constructed patch graph of N nodes with nodes $v_i \in \mathcal{V}$ and edges $e_{ij} = (v_i, v_j) \in \mathcal{E}$. Here, each patch is treated as a node and the edges in $\mathcal{E}$ are used to represent the relations between patches. And $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix where each element represents a pairwise relations of patches.

Motivated by [52], we represent the pairwise relations between every two patches in the graph as follows:

$$e\left(x_i, x_j\right) = \phi\left(x_i\right)^T \phi\left(x_j\right), \tag{2}$$

where ϕ represents a symmetrical transformation of the original patch features. More specifically, ϕ can be represented as $\phi = \mathbf{w}x$. The parameter $\mathbf{w}$ is a $d \times d$ dimension weight which is learnable via back propagation. By adding such transformation, it allows us to adaptively select and learn the correlations of different patches within a frame or across different frames.

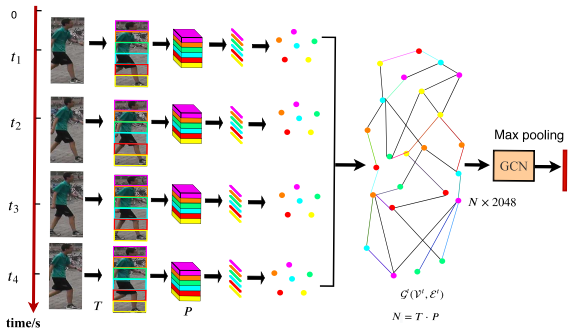

Figure 3. Illustration of the temporal GCN branch. The different color borders mean different patches. In this branch, we horizontally partition each feature map into P patches, then we can totally get $T \cdot P$ patches for a video sampling T frames. These patches are used as the nodes of graph. We can get the graph representation of the video, which be denoted as $\mathcal{G}^t(\mathcal{V}^t, \mathcal{E}^t)$. Then, we perform graph convolution operation on the graph. Finally, we use max pooling on the output of GCN to get final feature.

Then, the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ associated to $\mathcal{G}$ can be constructed, which is the key component for GCN because each element $\mathbf{A}_{ij}$ reflects the relations of node x_i and node x_j. However, considering the following two points, 1) for each row of the affinity matrix, the sum of all the edge values (i.e. the edges connect to patch i) should be 1; 2) each element of the adjacency matrix should be non-negative, the coefficient should be in the range of (0,1); we perform normalization operation on each row of the adjacent matrix $\mathbf{A}$ by

$$\mathbf{A}_{(i,j)} = \frac{e^2(x_i, x_j)}{\sum_{j=1}^{N} e^2(x_i, x_j)}. \tag{3}$$

Following Kipf and Welling [20], let $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n$ represents the self-loop adjacency matrix and $\mathbf{I_n} \in \mathbb{R}^{N \times N}$ is the identity matrix, we can use a re-normalization trick to approximate the graph-Laplacian:

$$\widehat{\mathbf{A}} = \widetilde{\mathbf{D}}^{-\frac{1}{2}} \widetilde{\mathbf{A}} \widetilde{\mathbf{D}}^{-\frac{1}{2}}, \tag{4}$$

where $\widetilde{\mathbf{D}}_{(i,i)} = \sum_j \widetilde{\mathbf{A}}_{(i,j)}$. Finally, we can obtain the corresponding adjacent matrix $\widehat{\mathbf{A}}$ for the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ so that we can further model the structural and temporal relations of patches.

3.3. Temporal GCN Module

As we have mentioned in Section 1, the patches of different frame in video can provide complementary information for alleviating the problem causing by occlusion and noise. In our proposed model, the Temporal GCN module (TGCN) is designed to capture the temporal dynamics relationships between patches across the different frames.

As shown in Figure 3, each video has N patches then we use all patches to construct the temporal graph $\mathcal{G}^t(\mathcal{V}^t, \mathcal{E}^t)$, where $\mathcal{V}^t = \{x_1, x_2, \ldots, x_N\}$, and the corresponding adjacent matrix $\widehat{\mathbf{A}}^t$ by using Equation (2) (3) (4).

For the temporal branch, given the adjacent matrix $\widehat{\mathbf{A}}^t$, we apply the GCN to capture the temporal relations of the patches of the whole video. We build M-layer graph convolutions in our implementation. Specifically for the m-th layer ($1 \leq m \leq M$), the graph convolution is implemented by

$$\mathbf{X}^m = \widehat{\mathbf{A}}^t \mathbf{X}^{m-1} \mathbf{W}^m. \tag{5}$$

where $\mathbf{X}^{(m)} \in \mathbb{R}^{N \times d_m}$ are the hidden features for all patches at layer m and d_m is the dimension of feature; $\mathbf{X}^0 \in \mathbb{R}^{N \times d}$ is the initial patches features obtained by CNN backbone; $\mathbf{W}^{(m)} \in \mathbb{R}^{d_m \times d_m}$ is the parameter matrix to be learned. After each layer of graph convolution, a Layer Normalization [2] layer and a LeakyReLU(with negative input slope α = 0.1) are appended. In addition, our experiments find it will be more effective and stable by using shortcut connection as [15],

$$\mathbf{X}^m := \mathbf{X}^m + \mathbf{X}^{m-1}, 2 \leq m \leq M. \tag{6}$$

After graph convolution, the output of temporal GCN module is $\mathbf{X}^M \in \mathbb{R}^{N^t \times d_M}$ for each video. Finally, we use max pooling operation on $\mathbf{X}^M$. Therefore, for each video, we can obtain its temporal GCN feature : $f^t \in \mathbb{R}^{1 \times d_M}$, where d_M is set to 2048 in our experiments.

3.4. Structural GCN Module

One of the most challenging difficulties of image-based person Re-ID is how to distinguish visually similar identities, and most of the image-based methods can only rely on extracting fine-grained appearance features. However, in video-based person Re-ID, the structural information (e.g. shape information) of the same identity will be more complete and precise because each video has many frames which may cover more views and poses. Thus the structural information can provide extra discriminative information to enhance a Re-ID system.

As shown in Figure 4, the structural GCN module (SGCN) is different from the TGCN. In TGCN, we use all patches of different frames to construct the graph and it aims at capturing complementary information among patches across frames. While in SGCN, we firstly use GCN to model the spatial relations of different patches for each frame in a video (i.e. each frame has a GCN). Then, we fuse the GCN features of frames in the video to get the intrinsic structural feature in the video.

Specifically, given a video with T frames, the GCN of i-th frame is represented as $\mathcal{G}_i^s(\mathcal{V}_i^s, \mathcal{E}_i^s)$, where $\mathcal{V}_i^s = \{x_{i,1}, x_{i,2}, \ldots, x_{i,P}\}$ (note that the subscript i represents the i-th frame and each frame is divided into P patches).

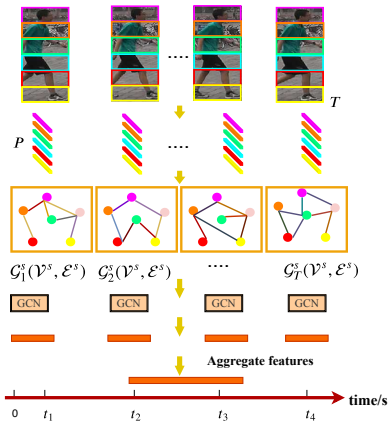

Figure 4. Illustration of the spatial GCN branch. We independently exploit the relations of the patches of each frame to capture the structural information from video sequences. We aggregate all the output features of GCNs to form the structural feature of the video.

Similar to TGCN, we use Equation (2) (3) (4) to obtain the corresponding adjacent matrix $\widehat{\mathbf{A}}_i^s$ for each $\mathcal{G}_i^s(\mathcal{V}^s, \mathcal{E}^s)$; then we build a K-layer graph convolutions for i-th frame. For the k-th ($1 \leq k \leq K$) graph convolution layer, the detailed operation can be written as

$$\mathbf{X}_i^k = \widehat{\mathbf{A}}_i^s \mathbf{X}_i^{k-1} \mathbf{W}_i^k, \tag{7}$$

where $\mathbf{W}_i^k \in \mathbb{R}^{d_k \times d_k}$, and d_k is the dimension of feature. To reduce the dimension of each sub GCN feature, the final output of GCN is $X_i^K \in \mathbb{R}^{P \times 256}$; then, we use max pooling operation so that dimension of the feature of each frame is 256. Finally, the features of the video are concatenated and the final feature is denoted as f^s.

3.5. Overview of Our Model and Loss Functions

As shown in Figure 2, our proposed model consists of a global branch, a temporal branch, and a spatial branch. The global branch extracts the global appearance feature for each video. The temporal branch with TGCN models the temporal relations of patches across different frames for learning temporal information, which can provide complementary information for other patches. The spatial branch with SGCN is designed for modeling the spatial relations for each frame to extract the structural information.

We use the batch hard triplet loss function [16] and the softmax cross-entropy loss function to train the networks. As shown in Figure 2, the two loss formulas are denoted as $L_{triplet}$ and $L_{softmax}$ respectively.

Specifically, in our experiments, we individually compute each triplet loss of three type features. Thus the final triplet loss can be represented as:

$$L'_{triplet} = L^{global}_{triplet} + L^{t}_{triplet} + L^{s}_{triplet}. \tag{8}$$

For the softmax cross-entropy loss function $L_{softmax}$, in our experiments, we concatenate the three type features f^{global},f^t,f^s as the final feature, which can be written as $f^{all} = [f^{global}, f^t, f^s]$, where $[\cdot]$ means concatenation. Finally, we use the feature f^{all} to compute the softmax cross-entropy loss.

Thus the total loss L_{total} is the combination of these two losses as follows:

$$L_{total} = L_{softmax} + L'_{triplet}. \tag{9}$$

4. Experiments

4.1. Datasets and Evaluation Protocols

Datasets. We evaluate our proposed model on two large-scale video-based person Re-ID datasets: DukeMTMC-VideoReID [58, 53] and MARS [71]. MARS is the largest video-based person re-identification benchmark dataset with 17,503 sequences of 1,261 identities and 3,248 distractor sequences. The training set contains 625 identities and the testing set contains 636 identities. DukeMTMC-VideoReID dataset is another large-scale benchmark dataset with 4,832 tracklets of 1,812 identities for video-based person Re-ID. It is derived from the DukeMTMC dataset [43]. The dataset is divided into 408, 702 and 702 identities for distraction, training, and testing, respectively. The bounding boxes are annotated manually.

Evaluation protocols. In our experiments, we adopt the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) to evaluate the performance of our proposed method.

4.2. Implementation Details

We use ResNet50 [15] as our backbone network which is pre-trained on ImageNet [9] firstly. The last stride of ResNet50 is set to be 1. We adopt a restricted random sampling strategy [26] to randomly sample T = 8 frames from every video. Each image is resized to 256×128 with random horizontal flips [32] for data augmentation. We train our network for 800 epochs in total, with an initial learning rate of 0.0003 and decayed it by 10 every 200 epochs. Adam [19] is chosen to optimize the networks. Following [16], we sample 16 identities, each with 4 tracklets, to form a batch of size $16 \times 4 \times 8 = 512$ images. For the parameters of GCN modules, the number of GCN layers in TGCN M is 3, the number of GCN layers in SGCN K is 2, and the number of patches P is 4.

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

To validate the effectiveness of our proposed method on the video-based person Re-ID problem, we compare our proposed method with several recent state-of-the-art methods on MARS and DukeMTMC-VideoReID. The results of the comparisons are presented in Table 1 and Table 2. We

Table 1. Performance (%) comparison with related works on MARS.

Method	mAP	rank1	rank5	rank20
BoW+kissme [71]	15.50	30.60	46.20	59.20
IDE+XQDA [71]	47.60	65.30	82.00	89.00
SeeForest [75]	50.70	70.60	90.00	97.60
QAN [33]	51.70	73.70	84.90	91.60
DCF [22]	56.05	71.77	86.57	93.08
TriNet [16]	67.70	79.80	91.36	-
MCA [47]	71.17	77.17	-	-
DRSA [26]	65.80	82.30	-	-
DuATM [46]	67.73	81.16	92.47	-
MGCAM [47]	71.17	77.17	-	-
PBR [48]	75.90	84.70	92.80	95.00
CSA [4]	76.10	86.30	94.70	98.20
STMP [34]	72.70	84.40	93.20	96.30
M3D [24]	74.06	84.39	93.84	97.74
STA [11]	80.80	86.30	95.70	98.10
GLTR [23]	78.47	87.02	95.76	98.23
Wu et al. [57]	81.1	89.8	96.1	97.6
VRSTC [17]	82.3	88.5	96.5	97.4
Zhao et al. [70]	78.2	87.0	95.4	**98.7**
STE-NVAN [32]	81.2	88.9	-	-
STGCN(Ours)	**83.70**	**89.95**	**96.41**	98.28

Table 2. Performance (%) comparison with related works on DukeMTMC-VideoReID.

Method	mAP	rank1	rank5	rank20
EUG [58]	78.3	83.6	94.6	97.6
ETAP-Net [58]	78.34	83.62	94.59	97.58
STE-NVAN [32]	93.5	95.2	-	-
VRSTC [17]	93.5	95.0	99.1	-
STA [11]	94.90	96.20	99.30	99.60
GLTR [23]	93.74	96.29	**99.30**	99.71
Wu et al. [57]	94.2	96.7	99.2	99.7
STGCN(Ours)	**95.70**	**97.29**	99.29	**99.72**

can see that our proposed method achieves the best results on rank-1 accuracy and mAP on both datasets.

Specifically, existing attention-based methods (including STA [11], GLTR [23]) process different regions and frames independently and they do not fully consider the intrinsic relations between patches. Thus they may miss some discriminative cues for Re-ID. Zhao *et al*. [70] requires extra attribute labels, which limits its application. As for M3D [24], the 3D convolutional operation is computationally expensive and sensitive to spatial misalignment. Particularly, compared with other graph-based method [57], our proposed method achieves better results on both datasets. The main reason can fall into two aspects: 1) the pose estimation in [57] is separate from the whole framework which may cause sub-optimal result and the pose alignment is sensitive to the quality of pose estimation; 2) it does not explicitly model the spatial relations of body parts for each frame, ignoring the structural information of intra-frame.

In summary, compared to existing methods, our proposed method jointly considers the potential relations of different parts of the body within a frame and across different frames which can provide more discriminative and robust information, and is able to perform end-to-end training. These experimental results validate the superiority of our method.

Table 3. The performance (%) of individual components in our proposed method. "SGCN+global" means we only used spatial and global branches during training and testing, and similar for "TGCN+global". For the baseline model, we remove the temporal branch and the spatial branch. "Ensemble" means we combine "SGCN+global" and "TGCN+global" by using score sum.

Dataset	MARS		DukeMTMC	
Method	mAP	rank1	mAP	rank1
Baseline (only global branch)	80.76	88.74	94.08	96.01
TGCN+global	81.97	89.70	95.12	96.87
SGCN+global	82.17	89.80	94.55	96.44
Ensemble	82.67	89.55	94.64	96.15
STGCN (SGCN+TGCN+global)	**83.70**	**89.95**	**95.70**	**97.29**

4.4. Ablation Study

4.4.1 The Impact of Two GCN Modules

To verify the impact of the spatial branch and the temporal branch separately, we train the baseline model, "SGCN+global", and "TGCN+global" under the same experiment setting of STGCN, respectively. "SGCN+global" means we only use spatial and global branches during training and testing, and similar for "TGCN+global". For the baseline model, we remove the temporal branch and the spatial branch. "Ensemble" means we combine "SGCN+global" and "TGCN+global" by using score sum. The experimental results are reported in Table 3.

In Table 3, the performance of "SGCN+global" and "TGCN+global" are higher than the performance of the baseline model, which verifies the effectiveness of each GCN module. Because the baseline model (i.e. only global branch) does not model the relations of patches, which contains more discriminative and robust information. We can see that STGCN and "Ensemble" achieve better results compared to "SGCN+global" and "TGCN+global", which means the temporal relations and structural relations are complementary. Furthermore, by comparing STGCN and "Ensemble", we also can conclude that STGCN is not only simply superimposing a single feature but benefits from jointly modeling temporal and structural relations.

4.4.2 The Impact of Graph Convolution

To verify the effectiveness of graph convolution, we replace the layers in GCN with fully-connected layers then performe training and testing under the same experimental setting. Specifically, considering one-layer GCN, the Equation (5) or (7) can be written as $\mathbf{Y} = \mathbf{AXW}$, where $\mathbf{A}$ is adjacent matrix, $\mathbf{X}$ is input, $\mathbf{W}$ is the parameters matrix of

Table 4. Performance (%) of our proposed model with graph convolution network or fully connected network. For the baseline model, we remove the temporal branch and the spatial branch. "SGCN+global" means we only use spatial and global branches during training and testing. "TFCN+SGCN+global" means we combine the global, spatial, and temporal branches during training and testing. And the others are similar.

Dataset	MARS		DukeMTMC	
Method	mAP	rank1	mAP	rank1
Baseline (only global branch)	80.76	88.74	94.08	96.01
TFCN+global	80.62	88.28	94.34	96.15
TGCN+global	**81.97**	**89.70**	**95.12**	**96.87**
SFCN+global	81.10	89.39	94.28	95.30
SGCN+global	**82.17**	**89.80**	**94.55**	**96.44**
TFCN+SGCN+global	81.51	89.24	95.62	97.15
TGCN+SFCN+global	82.65	89.90	95.24	96.30
TFCN+SFCN+global	82.27	88.94	95.03	96.15
SGCN+TGCN+global (STGCN)	**83.70**	**89.95**	**95.70**	**97.29**

the GCN layer. On the other hand, the formulation of the fully-connected network (FCN) can be written $\mathbf{Y} = \mathbf{XW}$, where $\mathbf{X}$ is input and $\mathbf{W}$ is the parameters matrix of the FCN layer. Compared with GCN, the FCN can be viewed as removing the adjacent matrix $\mathbf{A}$.

Thus, to evaluate the impact of the graph convolution, we replace the GCN layers with fully-connected layers for each GCN Module by removing all the adjacent matrices in Equation (5) and (7). We have following variants of our model, including: (1) "TFCN" means the GCN layers in TGCN is replaced by fully-connected layers; (2) "SFCN" means the GCN layers in SGCN is replaced by fully-connected layers; and (3) "TFCN + SFCN" means the GCN layers in TGCN and SGCN are both replaced. The experimental results are shown in Table 4.

As shown in Table 4, the performance of the models with the GCN module is significantly higher than the models without the GCN module, and the effect of fully-connected layers is limited or could be detrimental. This is because the fully-connected layers cannot model the relations of different patches, such methods cannot further mine the potential information in a video. Thus, the graph convolution operation and the modeling temporal and structural relations is necessary.

4.5. Visualization

Visualization of class activation maps. We visualise the class activation maps (CAMs) in Figure 5 by using Grad-CAM [44]. We can observe that the class activation maps of different frames of our proposed method have higher activation in the same discriminative area. Meanwhile, it is not difficult to find that our proposed method can focus on more discriminative cues by leveraging the spatial and temporal relations of patches.

Retrieval results analysis. As shown in Figure 6, we visualise the retrieval results of the same person. We can see that the top 5 results of our proposed method are all matching. However, the Rank-4 and Rank-5 results of the baseline model are disturbed by the samples of other identities with similar appearance or occlusion. Thus, the retrieval results prove our proposed method indeed alleviate the problem of similar appearances of different identities and occlusion problem.

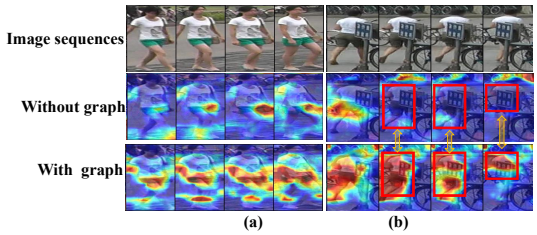

Figure 5. The visualization of the class activation maps (CAMs). The first row is the original image sequences from MARS. The second row is the class activation maps of the baseline model. The third row is the class activation maps of our proposed model.

4.6. Further Analysis

4.6.1 The Number of GCN Layers in GCN Module

In our proposed model, the number of GCN layers in TGCN and SGCN are denoted as M and K, respectively. We carry out experiments to investigate the effect of the number of GCN layers by changing one of the GCN modules while freezing the other one.

The impact of the number of GCN layers in TGCN. In this experiment, we fix the number of GCN layers in SGCN (i.e., $K = 2$) then evaluate the performance of our model when $M = 2, 3, 4, 5, 6$. From Figure 7 (a), we can see that the best Rank-1 is 90.35% when $M = 4$, and the best mAP is 83.70% then $M = 3$. Whether $M = 3$ or $M = 4$, the results outperform the state-of-the-art methods and baseline by a large margin.

The impact of the number of GCN layers in SGCN. Similarly, we fix the number of GCN layers in SGCN (i.e., $M = 4$) then evaluate the performance of our model when $K = 1, 2, 3, 4$. As shown in Figure 7 (b). When $K = 2$, the model achieves the best performance.

As shown in Figure 7, the performance of STGCN mostly higher than the baseline model (i.e., 80.76%/88.74% in mAP/Rank-1), although the number of GCN layers will affect the performance of the model. We also can observe that if two GCN modules are too shallow or deep, the effect of GCN will decrease. The graph convolution of the GCN can be simply view as a special form of Laplacian smoothing, which mixes the features of a vertex and its nearby neighbors. A shallow GCN cannot effectively propagate the node information to the entire data graph. But when the GCN is too deep, it also brings potential concerns of

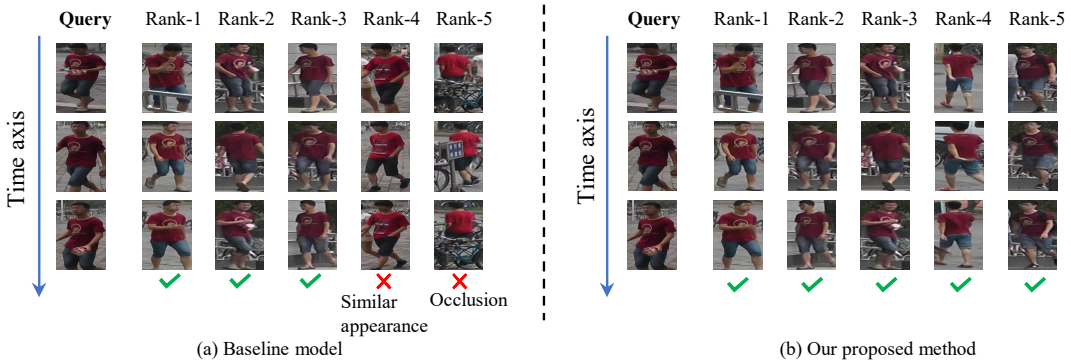

Figure 6. (a) and (b) are the top 5 retrieval results of the baseline model and our proposed method in the MARS dataset, respectively. The query and gallery both are image sequences. **Best viewed in color.**

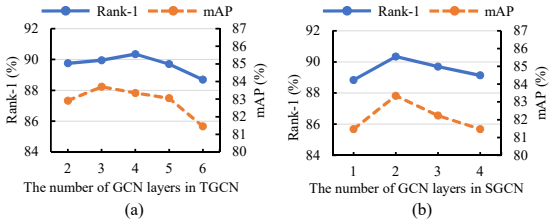

Figure 7. (a) Analysis on the number of GCN layers in TGCN (b) Analysis on the number of GCN layers in SGCN. We carry out these experiments on the MARS dataset.

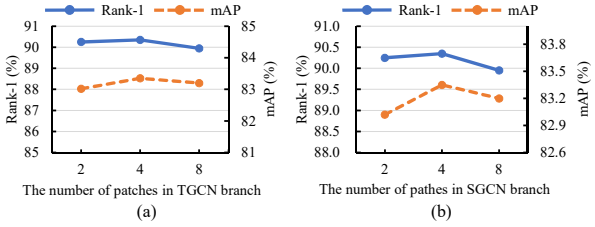

Figure 8. (a) Analysis on the number of patches in TGCN. (b) Analysis on the number of patches in SGCN. We carry out these experiments on the MARS dataset.

over-smoothing [25].

4.6.2 Analysis on the number of patches in GCN module

The number of nodes in the graph (i.e., the number of patches) is another key parameter of GCN. For convenience, we denote the number of patches of each frame in TGCN and SGCN are P^t and P^s, respectively.

The impact of the number of patches in TGCN. In this experiment, we fix $P^s = 4$ and evaluate the results when $P^t = 2, 4, 8$. From the Figure 8 (a), we can see the model achieves the best performance when P^t=4.

The impact of the number of patches in SGCN. Similarly, when we analyze the effect of P^s, we fix $P^t = 4$. As shown in Figure 8 (b), the model has the best performance when p^s=4.

From Figure 8, we can observe that our proposed model is robust to the number of nodes in the graph to some extent and the performance of these experiments outperforms the baseline model significantly. However, if the number of patches of each frame is too large or small, the performance will decrease. Because when the number of patches increases, the patches will become smaller, which cannot contain enough information. Conversely, when the number of patches is too small, the patches might ignore subtle but discriminative cues.

5. Conclusions

In this paper, we demonstrate the effectiveness of leveraging the temporal relations of patches for alleviating occlusion problem and the spatial relations of patches for distinguishing the ambiguity samples with similar appearance. Specifically, we propose a novel Spatial-Temporal Graph Convolutional Network (STGCN), which contains two core GCN branches. The spatial branch learns the structural information of human body by modeling relations of patches of each frame. The temporal branch can alleviate occlusion problem by modeling the temporal relations of patches across the different frames. Furthermore, we integrate the spatial branch and the temporal branch into a unified framework and jointly optimize the model. Extensive experiments validate the effectiveness of our proposed method.

6. Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC(U1911401,U1811461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong NSF Project (No. 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03).

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015. 1, 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2990–2999, 2017. 2

[4] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018. 1, 2, 6

[5] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*, 2018. 2

[6] Ying-Cong Chen, Wei-Shi Zheng, Jian-Huang Lai, and Pong C Yuen. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE transactions on circuits and systems for video technology*, 27(8):1661–1675, 2016. 2

[7] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 2

[8] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1983–1991, 2017. 1, 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[10] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010. 1, 2

[11] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the Association for the Advancement of Artificial Intelligence*. 2019. 1, 2, 6

[12] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018. 2

[13] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008. 1, 2

[14] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5, 6

[17] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019. 6

[18] Wenbing Huang, Tong Zhang, Yu Rong, and Junzhou Huang. Adaptive sampling towards fast graph representation learning. In *Advances in Neural Information Processing Systems*, pages 4558–4567, 2018. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2, 4

[21] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012. 2

[22] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 6

[23] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3958–3967, 2019. 2, 6

[24] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, 2019. 2, 6

[25] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 8

[26] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018. 2, 5, 6

[27] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 1, 2

[28] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3610–3617, 2013. 2

[29] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 1, 2

[30] Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5771–5780, 2017. 1, 2

[31] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *European Conference on Computer Vision*, pages 391–401. Springer, 2012. 1, 2

[32] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019. 5, 6

[33] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799, 2017. 2, 6

[34] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019. 6

[35] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014. 2

[36] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016. 1, 2

[37] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016. 2

[38] A Mignon and F Jurie. A new approach for distance learning from sparse pairwise constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2

[39] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015. 2

[40] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3318–3325, 2013. 2

[41] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010. 2

[42] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

[43] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 5

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[45] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 486–504, 2018. 2

[46] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 6

[47] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 6

[48] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 6

[49] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 1, 2

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[51] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016. 1, 2

[52] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018. 2, 3

[53] Xiaogang Wang and Rui Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*, pages 351–370. Springer, 2014. 5

[54] Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(8):1447–1460, 2015. 2

[55] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification.

Journal of Machine Learning Research, 10(Feb):207–244, 2009. 2

[56] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng. An enhanced deep feature representation for person re-identification. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016. 1, 2

[57] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, and Qi Tian. Adaptive graph representation learning for video person re-identification. *arXiv preprint arXiv:1909.02240*, 2019. 2, 6

[58] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi 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*, pages 5177–5186, 2018. 5, 6

[59] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016. 1, 2

[60] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014. 1, 2

[61] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 4733–4742, 2017. 1, 2

[62] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[63] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016. 2

[64] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2158–2167, 2019. 2

[65] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2016. 2

[66] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 2

[67] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1239–1248, 2016. 2

[68] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 1, 2

[69] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017. 1, 2

[70] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xiansheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4913–4922, 2019. 6

[71] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 2, 5, 6

[72] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, 2012. 2

[73] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):591–606, 2015. 2

[74] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015. 1, 2

[75] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017. 2, 6