Spatial-Temporal Correlation and Topology Learning for Person Re-Identification in Videos

Jiawei Liu, Zheng-Jun Zha*, Wei Wu, Kecheng Zheng, Qibin Sun
University of Science and Technology of China, China
{jwliu6,zhazj,qibinsun}@ustc.edu.cn, {wuvy,zkcys001}@mail.ustc.edu.cn

Abstract

Video-based person re-identification aims to match pedestrians from video sequences across non-overlapping camera views. The key factor for video person re-identification is to effectively exploit both spatial and temporal clues from video sequences. In this work, we propose a novel Spatial-Temporal Correlation and Topology Learning framework (CTL) to pursue discriminative and robust representation by modeling cross-scale spatial-temporal correlation. Specifically, CTL utilizes a CNN backbone and a key-points estimator to extract semantic local features from human body at multiple granularities as graph nodes. It explores a context-reinforced topology to construct multi-scale graphs by considering both global contextual information and physical connections of human body. Moreover, a 3D graph convolution and a cross-scale graph convolution are designed, which facilitate direct cross-spacetime and cross-scale information propagation for capturing hierarchical spatial-temporal dependencies and structural information. By jointly performing the two convolutions, CTL effectively mines comprehensive clues that are complementary with appearance information to enhance representational capacity. Extensive experiments on two video benchmarks have demonstrated the effectiveness of the proposed method and the state-of-the-art performance.

1. Introduction

Person re-identification (Re-ID) is an important technology to retrieve a person-of-interest across non-overlapping cameras. It has drawn increasing attention during the past few years, owing to its broad application in many realistic scenarios, such as video surveillance [11, 47] and behavior analysis [43] *etc*. However, this task remains challenging due to the variations in illumination, viewpoint and pose, as well as the influence of background clutter and occlusion.

Existing person Re-ID approaches are mainly divided into two categories: image-based methods [39, 33, 41, 25] and video-based methods [36, 26, 22]. The former exploits static images without temporal information to retrieve pedestrians. It has achieved impressive advances with the surge of deep learning technique in recent years [18]. However, image-based person Re-ID heavily relies on the quality of static images, which are sensitive to noise, occlusion and viewpoint variation, *etc*. Different from static images with limited content, video sequences contain rich spatial-temporal information across a long span of time, which can provide clean and informative clues against these problem [22, 8]. Thus, video-based person Re-ID has the potential to solve the restrictions in image-based person Re-ID.

A typical video-based person Re-ID pipeline extracts and aggregates spatial and temporal clues from video sequences to generate discriminative representations. Some preliminary methods [29, 10, 13, 52] extract appearance features from each frame independently, and aggregate them into video-level representation by temporal pooling layer or recurrent neural network (RNN). In presence of partial occlusions, inaccurate detection and viewpoint vari-

Figure 1. Three example video sequences on MARS and iLIDS-VID datasets with partial occlusions, inaccurate detection and viewpoint variation.

* Corresponding author.

使用姿态估计提取对齐的局部特征

ation, the learned features are often corrupted, result in significant performance degradation. Figure 1 illustrates some video sequences of pedestrians on MARS [52] and iLIDS-VID [42] datasets with these issues. Recent works attempt to address them by dividing video frames into horizontal rigid stripes [8, 6, 43] or utilizing attention mechanism [22, 21, 36, 31, 14] to discover distinctive partial regions for extracting local appearance features. However, much background noise is blended in their located partial regions, thus they can not learn precise aligned part features from videos [53]. Considering that, a few works [3, 19, 9, 51] employ pose estimation model [30] to adaptively locate key-points of pedestrians for extracting aligned part features. However, drastic viewpoint and pose variations as well as occlusion within videos affect the reliability of pose estimation model. Meanwhile, these methods only extract local features with fixed semantics from one-granularity partition, which can not cover all discriminative clues. Further, all the aforementioned methods only model the temporal relation across different frames, while neglecting complicated spatial-temporal dependencies and structural information of different body parts within a frame or across frames, restricting the capability of pedestrian representation.

In this work, we propose a novel Spatial-Temporal Correlation and Topology Learning framework (CTL) for video-based person re-identification, which pursues discriminative and robust representations. CTL extracts local features at multi-granularity levels to capture diverse discriminative semantics and alleviate unstable pose estimation results, and learns the potential cross-scale spatial-temporal dependencies and structure information among body parts for enhancing feature representation. Specifically, CTL employs a CNN backbone and a key-points estimator to extract semantic part features from human body at three granularities as graph nodes. It then explores a context-skeleton enriched topology to construct multi-scale graphs by considering both global contextual information and physical connections of human body, which effectively models the intrinsic spatial-temporal linkages between nodes. Moreover, a 3D graph convolution and a cross-scale graph convolution are designed for these multi-scale graphs, which facilitate direct cross-spacetime and cross-scale information propagation for capturing hierarchical spatial-temporal dependencies and structural information. By jointly performing the two convolutions, CTL effectively mines comprehensive and discriminative clues that are complementary with appearance information to enrich representation. Extensive experiments on two video datasets, *i.e.*, MARS and iLIDS-VID, have demonstrated the effectiveness of the proposed approach.

Although graph modeling has been explored in person Re-ID, most of them only construct a graph on image-level [33, 1, 45, 23] without considering temporal relation. A few of preliminary works [43, 44, 46] extend graph modeling to video person Re-ID. However, they neglect the spatial structural information within each frame [43, 44], or simply utilize factorized spatial and temporal graph modeling [46], failing to capture complex spatial-temporal relation. Further, all of them essentially belong to a local method. They utilize pair-wise feature affinity to measure the linkage between two nodes, while ignoring the impact of global contextual information from all other nodes, which is significance for learning reliable and useful graph topology.

The main contributions of this paper are as following: (1) We propose a novel Spatial-Temporal Correlation and Topology Learning framework (CTL) for person re-identification in videos. (2) We learn a context-reinforced topology to construct multi-scale graphs by considering both global contextual information and physical connections of human body. (3) We develop a 3D graph convolution and a cross-scale graph convolution to model high-oder spatial-temporal dependencies and structural information.

2. Related Work

Image-based Person Re-ID. It is extensively explored in the literature. Existing methods mainly focus on three categories: designing discriminative hand-crafted descriptors [2], robust distance metric learning [24, 50] or deep learning technique [27, 39, 18, 17, 16]. For example, Chen *et al.* [5] introduced a cascaded feature suppression mechanism that mines all potential salient features stage-by-stage and integrates these discriminative salience features with the global feature, producing the final pedestrian feature.

Video-based Person Re-ID. Compared with image-based person Re-ID, video-based person Re-ID provides richer spatial-temporal clues and is promising for precise retrieval [37, 15, 48]. Some existing works [29, 10, 13] formulate video-based person Re-ID as an extension of image-based person Re-ID. They extract appearance representation from each frame, and aggregates the representations of all frames by using temporal pooling layer or RNN. For example, McLaughlin *et al.* [29] proposed a siamese network, which captures features from each video, and then employs a recurrent layer and a temporal pooling layer to abstract video-level feature. In order to learn robust representation against partial occlusions, inaccurate detection and pose variation, rigid stripe partition [8, 6, 43] and attention mechanism [22, 21, 36, 31] methods attract more attention recently. For example, Subramaniam *et al.* [36] formulated a Co-segmentation Activation Module to enhance common abstract features and suppress background features by jointly exploring common features across frames. Moreover, a few works [3, 19, 9] utilize pose estimation model to adaptively locate key-points of human body and learn aligned semantic features. For example, Jones *et al.* [19] proposed a pose-guided alignment framework, which mim-

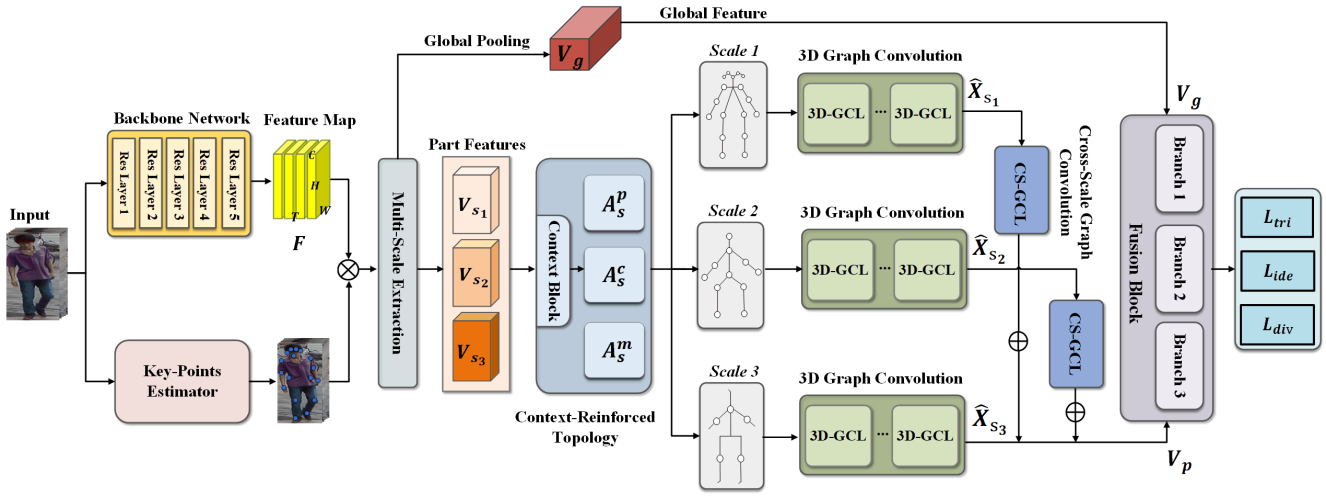

Figure 2. The overall architecture of the proposed CTL. It consists of a backbone network with a key-points estimator, a context block, multiple 3D graph convolution layers (3D-GCLs), multiple cross-scale graph convolution layers (CS-GCLs) and a fusion block.

icked the top-down attention of the human visual cortex to learn aligned features.

Graph Learning. Graphs are typically utilized to model relationships between nodes. Graph convolutional network (GCN) and its variant models [20] have achieved great success in many computer vision tasks, *e.g.*, object detection [35], multi-label image recognition [7] and skeleton-based action recognition [49, 28]. Similarity, some methods also apply GCN to person Re-ID. Most of them [33, 1, 45, 23] build the graph models on image-level by considering the relations among images, which neglect the beneficial temporal information. In addition, a few of recent works [43, 44, 46] extent GCN to video person Re-ID by exploring spatial and temporal relation, while they ignore the spatial structural information of body parts within each frame [43, 44] or only consider factorized spatial and temporal relation modeling [46]. For example, Yang *et al.* [46] proposed a Spatial-Temporal Graph Convolutional Network (STGCN) which includes two GCN branches. The spatial branch learns spatial relation of human body, and the temporal branch mines discriminative temporal relation from adjacent frames.

3. Method

To further enhance the capacity of representations, this work explicitly explores spatial-temporal features across multi-granularity levels. To this end, we propose a Spatial-Temporal Correlation and Topology Learning framework, which models high-order spatial-temporal correlation to learn comprehensive representation. The overall architecture is shown in Figure 2. It consists of a backbone network with a key-points estimator, a context block, multiple 3D graph convolution layers (3D-GCL), multiple cross-scale graph convolution layers (CS-GCL) and a fusion block.

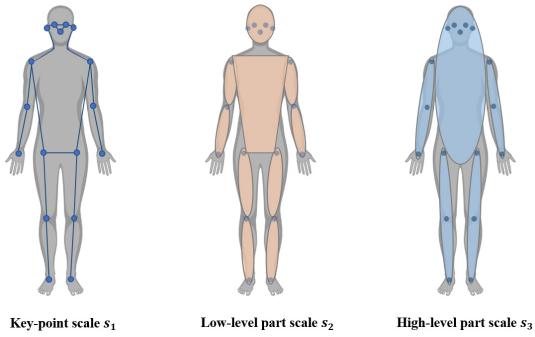

Figure 3. Three scales of body partition. In s_1, we adopt 17 key-points, In s_2 and s_3, we adopt 10 and 5 parts of body, respectively.

3.1. Multi-Scale Feature Extraction

Given a video sequence, we denote it as $\{\boldsymbol{I}_t\}_{t=1}^{T}$, where T is the sequence length. The backbone network takes each frame as input to extract the initial feature map $\boldsymbol{F} = \{\boldsymbol{F}_t | \boldsymbol{F}_t \in \mathbb{R}^{H\times W\times C}\}_{t=1}^{T}$, where H, W and C denote the height, width and channel size of the feature maps, respectively. The backbone network is based on ResNet-50 model [12]. As part-based representations have shown effectiveness for person Re-ID [39], we adopt a key-points estimator [38] to adaptively locate the key-points of human body, and extract aligned part features from these key-points against partial occlusions, misalignment and viewpoint variation. Although, key-points estimation models have obtained high accuracy, they remain suffering from unreliable performance under complex surveillance scenes, leading to inaccurate key-points location and their confidence. Thus, exploring multi-scale part features with their spatial-temporal correlation is particularly important, which can alleviate unreliable key-point estimation results and

capture diverse discriminative semantics.

Based on human nature, we divide human body at three granularities: the key-point scale (s_1), the low-level-part scale (s_2) and the high-level-part scale (s_3), as show in Figure 3. We merge spatially nearby key-points to each part in coarser scales based on human prior. The heat maps $\boldsymbol{m}$ of key-points are generated through the key-point estimator, which are then normalized with a softmax function. The group of semantic local feature $\boldsymbol{V}_{s1} \in \mathbb{R}^{T \times N_{s_1} \times C}$ for the granularity s_1 and the global feature $\boldsymbol{V}_g \in \mathbb{R}^{T \times C}$ are computed as following:

$$\begin{aligned} \boldsymbol{V}_{s1} &= \{\boldsymbol{v}_{s1}\} = g_{GAP}(\boldsymbol{F}_t \otimes \boldsymbol{m}_{s_1}^t) \\ \boldsymbol{V}_g &= g_{GAP}(\boldsymbol{F}) \end{aligned} \tag{1}$$

where $\otimes$ and g_{GAP} refer to outer product and global average pooling operations, respectively. N_s is the number of body parts (*17, 10 and 5 parts for* s_1, s_2, s_3, *respectively*). The part features $\boldsymbol{V}_{s2}$ and $\boldsymbol{V}_{s3}$ for the low-level-part and high-level-part scales are computed by performing average pooling operation on the features $\boldsymbol{V}_{s1}$ of the key-points within each body part.

3.2. Context-Reinforced Topology Graph

In order to excavate spatial-temporal information from video frames, we employ advanced GCN to model hierarchical spatial-temporal dependencies and structural information. Let $\mathcal{G} = \{\mathcal{G}_s\}_{s \in \{s_1, s_2, s_3\}}$ be a set of constructed multi-scale graphs of one video frame, where each graph corresponds to a specific granularity level s. Specifically, $\mathcal{G}_s(\mathcal{V}_s, \mathcal{E}_s)$ includes N_s nodes $\boldsymbol{v}_i \in \mathcal{V}_s$ and a set of edges $\boldsymbol{e}_{ij} = (\boldsymbol{v}_i, \boldsymbol{v}_j) \in \mathcal{E}_s$. Each part of body within one video frame is viewed as a graph node and the edges represent the relationship between these body parts. The input node feature of frame t is denoted as $\boldsymbol{X}_s^t = \boldsymbol{V}_s^t \in \mathbb{R}^{N_s \times C}$. $\boldsymbol{A}_s \in \mathbb{R}^{N_s \times N_s}$ is the corresponding frame-level adjacency matrix, in which each element represents the linkage of two arbitrary nodes. The topology of the graph is actually decided by $\boldsymbol{A}_s$. Existing GCN-based Re-ID methods predict the relationship between two nodes independently by calculating pair-wise feature affinity, which ignore the impact of all other contextual nodes and only consider undirected dependency, restricting the capacity and expressiveness of the graph model.

Considering that, we explore a context-reinforced topology to construct graph, which simultaneously encodes contextual information along the node, temporal and feature dimensions, as well as physical structural information of human body. The context-reinforced adjacency matrix $\boldsymbol{A}_s$ consists of three components:

$$\boldsymbol{A}_s = \boldsymbol{A}_s^p + \boldsymbol{A}_s^m + \boldsymbol{A}_s^c \tag{2}$$

where $\boldsymbol{A}_s^p \in \{0, 1\}^{N_s \times N_s}$ denotes the physical connections of human body with rich structural information, which is

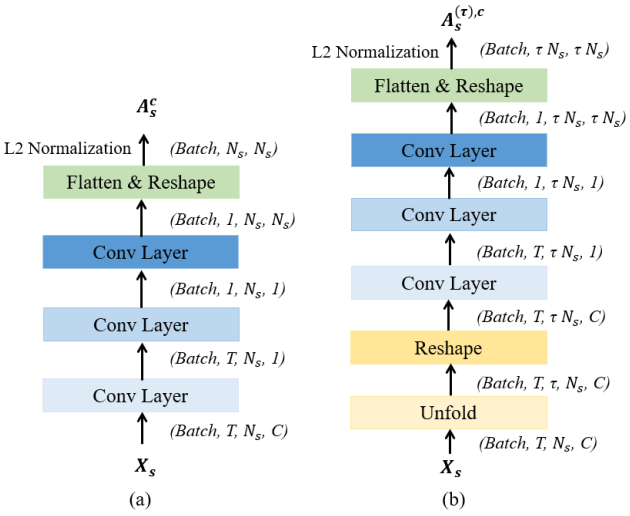

Figure 4. Detailed network structure of (a) the context block; (b) the advanced context block.

fixed during training. $\boldsymbol{A}_s^m$ denotes a mask adjacency matrix, which is utilized as the attention on the physical structure, inspired by [34]. $\boldsymbol{A}_s^m$ improves the flexibility and generality of static global graph structure $\boldsymbol{A}_s^p$, and is initialized with zeros and optimized together with other parameters during training. $\boldsymbol{A}_s^c$ is a data-dependent individual adjacency matrix, which incorporates the global contextual information of all nodes and learns a unique dynamic topology graph for each sample. $\boldsymbol{A}_s^c$ is learned by a context block, as shown in Figure 4(a). Given the node features $\{\boldsymbol{X}_s^t\}_{t=1}^T \in \mathbb{R}^{T \times N_s \times C}$, the context block firstly squeezes the feature and temporal dimensions of each node by two convolution layers with 1×1 kernel. Then, it utilizes an addition 1×1 convolution layer to transfer the N_s-dimension feature vector into the $N_s \times N_s$ adjacency matrix $\boldsymbol{A}_s^c$. Afterwards, $L2$ normalization operation is applied to each row of $\boldsymbol{A}_s^c$ for stable optimization. The context block adequately considers the influence of all other nodes when measuring the relationship between two arbitrary nodes.

3.3. 3D Graph Convolutional Layer

After obtaining the frame-level graphs for all frames, we design a 3D graph convolution to effectively propagate messages and update node features. 3D-GCL allows direct cross-spacetime information propagation for capturing complex spatial-temporal dependencies and structural information in a spatial-temporal graph, as shown in Figure 5(a). Concretely, 3D-GCL first uses a temporal sliding window with size of τ over the sequence of frame-level graphs. At each sliding step, a spatial-temporal subgraph $\mathcal{G}_s^{(\tau)} = (\mathcal{V}_s^{(\tau)}, \mathcal{E}_s^{(\tau)})$, in which $\mathcal{V}_s^{(\tau)} = \mathcal{V}_s^1 \cup ... \cup \mathcal{V}_s^\tau$ denote the union set of all nodes across τ video frames in this window. And the edge set $\mathcal{E}_s^{(\tau)}$ is represented by a block

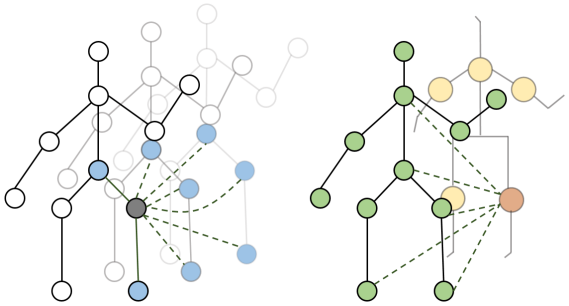

Figure 5. (a) Cross-spacetime information propagation by 3D graph convolution; (b) Cross-scale information propagation by cross-scale graph convolution.

adjacency matrix $\boldsymbol{A}_s^{(\tau)}$. It is computed as following:

$$\begin{aligned}\boldsymbol{A}_s^{(\tau)} &= \boldsymbol{A}_s^{(\tau),p} + \boldsymbol{A}_s^{(\tau),m} + \boldsymbol{A}_s^{(\tau),c} \\ &= \begin{bmatrix} [\boldsymbol{A}_s^{(\tau)}]_{1,1} & \cdots & [\boldsymbol{A}_s^{(\tau)}]_{1,\tau} \\ \vdots & \ddots & \vdots \\ [\boldsymbol{A}_s^{(\tau)}]_{\tau,1} & \cdots & [\boldsymbol{A}_s^{(\tau)}]_{\tau,\tau} \end{bmatrix} \in \mathbb{R}^{\tau N_s \times \tau N_s}\end{aligned} \tag{3}$$

where each submatrix $[\boldsymbol{A}_s^{(\tau)}]_{i,j}$ denotes the graph nodes of $\mathcal{V}_s^i$ are connected to themselves and their temporal neighboring nodes at frame j, by expanding the frame-level spatial connection (corresponding to $[\boldsymbol{A}_s^{(\tau)}]_{i,i}$) to the temporal domain. $\boldsymbol{A}_s^{(\tau)}$ is still composed of three parts. The block adjacency matrix $\boldsymbol{A}_s^{(\tau),p}$ is computed by tiling the static $\boldsymbol{A}_s^p$ in each block. $\boldsymbol{A}_s^{(\tau),m}$ is obtained as the same way. $\boldsymbol{A}_s^{(\tau),c}$ is learned by the advanced context block in Figure 4(b). Simultaneously, $\boldsymbol{X}_s^{(\tau)} \in \mathbb{R}^{T \times \tau N_s \times C}$ is obtained by employing the sliding temporal window over $\boldsymbol{X}^0 = \{\boldsymbol{X}_s^t\}_{t=1}^T \in \mathbb{R}^{T \times N_s \times C}$ with zero padding operation to build T windows, which is the input of 3D-GCL.

The 3D graph convolution for the t-th temporal window at l-th iteration is formulated as following:

$$[\boldsymbol{X}_s^{(\tau),l+1}]_t = \sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}} \hat{\boldsymbol{A}}_{s,t}^{(\tau)} \tilde{\boldsymbol{D}}^{-\frac{1}{2}} [\boldsymbol{X}_s^{(\tau),l}]_t \boldsymbol{W}^l) \tag{4}$$

where $(\tilde{\boldsymbol{D}})_{i,i} = \sum_j (\hat{\boldsymbol{A}}_{s,t}^{(\tau)})_{i,j}$ denotes diagonal node degree matrix [20], $\hat{\boldsymbol{A}}_{s,t}^{(\tau)} = \boldsymbol{A}_{s,t}^{(\tau)} + I_{\tau N_s}$ denotes the self-loop adjacency matrix. $\boldsymbol{I}_{\tau N_s}$ denotes an identity matrix. $\boldsymbol{W}^l$ refers to the learnable parameter and σ represents a non-linear activation function. After each 3D-GCL, a convolution layer followed with a batch normalization (BN) layer and a rectified linear units (ReLU) layer is employed to collapse the window dimension τ and output the updated node feature $\boldsymbol{X}^{l+1} \in \mathbb{R}^{T \times N_s \times C}$. In addition, shortcut connection $\boldsymbol{X}^{l+1} = \boldsymbol{X}^{l+1} + \boldsymbol{X}^l, 1 \leq l \leq L-1$ is adopted for effective and stable optimizing. The refined part features at three granularities $\hat{\boldsymbol{X}}_{s_1}$, $\hat{\boldsymbol{X}}_{s_2}$, $\hat{\boldsymbol{X}}_{s_3}$ are finally obtained by performing multiple 3D-GCLs.

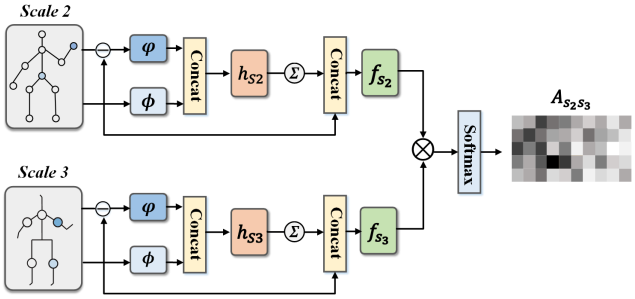

Figure 6. The inference of the cross-scale adjacent matrix.

3.4. Cross-Scale Graph Convolutional Layer

Multi-scale part features obtained from different partitions contain diverse discriminative semantics. To enable information diffusion across scales and learn comprehensive representation, we propose a cross-scale graph convolution, which propagates the informative clues of part features from one scale to another, as shown in Figure 5(b). The cross-scale topology graph is a directed graph that corresponds the nodes in one scale graph to the nodes in another scale graph. For simplicity, we elaborate a CS-GCL associated from s_2 to s_3. The adjacent matrix $\boldsymbol{A}_{s_2,s_3} \in \mathbb{R}^{N_{s_3} \times N_{s_2}}$ of the cross-scale graph predicts the cross-scale relationship. As shown in Figure 6, the dependency $(\boldsymbol{A}_{s_2,s_3})_{i,m}$ between i-th part in s_2 and m-th part in s_3 is computed as following:

$$\begin{aligned}\boldsymbol{p}_{i,s_2} &= \sum_{j=1}^{N_{s_2}} h_{s_2}([\phi(\boldsymbol{x}_{i,s_2}), \varphi(\boldsymbol{x}_{j,s_2} - \boldsymbol{x}_{i,s_2})]) \\ \boldsymbol{r}_{i,s_2} &= f_{s_2}([\boldsymbol{x}_{i,s_2}, \boldsymbol{p}_{i,s_2}]) \\ \boldsymbol{p}_{m,s_3} &= \sum_{j=1}^{N_{s_3}} h_{s_3}([\phi(\boldsymbol{x}_{m,s_3}), \varphi(\boldsymbol{x}_{j,s_3} - \boldsymbol{x}_{m,s_3})]) \\ \boldsymbol{r}_{m,s_3} &= f_{s_3}([\boldsymbol{x}_{m,s_3}, \boldsymbol{p}_{m,s_3}]) \\ (\boldsymbol{A}_{s_2,s_3})_{i,m} &= softmax(\boldsymbol{r}_{m,s_3}^\top \boldsymbol{r}_{i,s_2})\end{aligned} \tag{5}$$

where $\boldsymbol{x}_{i,s_2} \in \mathbb{R}^C$ denotes the i-th component of $\hat{\boldsymbol{X}}_{s_2}$ at one specific frame. h_{s_2}, f_{s_2}, ϕ, φ are the embedding functions implemented by a full connected layer with a BN layer and a ReLU layer. $\boldsymbol{p}_{i,s_2}$ and $\boldsymbol{p}_{m,s_3}$ aggregate the global relation information of all other part features to the i-th and the m-th components at the two scales. $\boldsymbol{r}_{i,s_2}$ and $\boldsymbol{r}_{m,s_3}$ are the augmented global relation features, which are then used to calculate the dependency $(\boldsymbol{A}_{s_2,s_3})_{i,m}$ by inner product operation and softmax function. Thus, $\boldsymbol{A}_{s_2,s_3}$ constructs the influence from the body in s_2 to each part in s_3.

Given the part feature $\hat{\boldsymbol{X}}_{s_2}$ at scale s_2, the cross-scale convolution for frame t is formulated as following:

$$[\hat{\boldsymbol{X}}_{s_{23}}]_t = \sigma(\boldsymbol{A}_{s_2,s_3}^t [\hat{\boldsymbol{X}}_{s_2}]_t \boldsymbol{W}_{s_{23}}) \tag{6}$$

where $\boldsymbol{W}_{s_{23}}$ denotes the parameter matrix, and $\hat{\boldsymbol{X}}_{s_{23}}$ is the transformed part feature. Such feature adaptively absorbs informative clues from the corresponding parts of body in s_2. Analogously, we also utilize another CS-GCL to transfer the part feature $\hat{\boldsymbol{X}}_{s_1}$ from s_1 to s_3, and produce the transformed part feature $\hat{\boldsymbol{X}}_{s_{13}}$. Finally, the comprehensive part feature $\boldsymbol{V}_p$ with three-granularity information is obtained as following:

$$\boldsymbol{V}_p = \hat{\boldsymbol{X}}_{s_3} + \alpha(\hat{\boldsymbol{X}}_{s_{13}} + \hat{\boldsymbol{X}}_{s_{23}}) \tag{7}$$

where α denotes the balance weight.

3.5. Model Optimizing

After obtaining the features $\boldsymbol{V}_p \in \mathbb{R}^{T\times N_{s_3}\times C}$ and $\boldsymbol{V}_g$, they are fed into a fusion block to further incorporate the global and local information, and are finally optimized by loss function. The fusion block consists of three branches. The first branch employs a temporal average pooling layer (g_{TAP}) for $\boldsymbol{V}_g$ to generate the feature vector $\boldsymbol{V}_f^g = g_{TAP}(\boldsymbol{V}_g)$. The second branch utilizes the function $g_{TAP}((\sum_{n=1}^{N_{s_3}}[\boldsymbol{V}_p]_{:,n,:}) + \boldsymbol{V}_g)$ to generate the feature vector $\boldsymbol{V}_f^a$. The third branch utilizes the function $g_{TAP}([g_c([\boldsymbol{V}_p]_{:,1,:}), ..., g_c([\boldsymbol{V}_p]_{:,N_{s_3},:})] + \boldsymbol{V}_g)$ to produce the feature vector $\boldsymbol{V}_f^c$, where g_c denotes a 1×1 convolution layer for reducing the dimension. The third branch implicitly promotes the channel-wise semantic alignment between the global and local features, which drives different channel of the global feature to focus on different body parts for improving the performance. Identification loss and triplet loss are the widely-used losses for person re-identification, we adopt triplet loss with hard mining strategy [46] and identification loss with label smoothing regularization [40] to optimize these three features $\boldsymbol{V}_f^g$, $\boldsymbol{V}_f^a$ and $\boldsymbol{V}_f^c$, respectively. The two losses are denoted as $\mathcal{L}_{tri}$ and $\mathcal{L}_{ide}$ respectively. Moreover, a diversity regularization loss is proposed to encourage the diversity of the local features and increase the discrimination of the final video representation. This loss is defined as following:

$$\mathcal{L}_{div} = \|\boldsymbol{V}_p\boldsymbol{V}_p^T - \boldsymbol{I}\|_F^2 \tag{8}$$

where $\|\cdot\|_F$ denotes Frobenius norm. $\boldsymbol{V}_p$ is applied with temporal average pooling and $L2$ normalization in advance for this loss. Therefore, the total loss $\mathcal{L}$ for CTL is the combination of the three losses:

$$\mathcal{L} = \lambda_1\cdot\mathcal{L}_{tri} + \lambda_2\cdot\mathcal{L}_{ide} + \lambda_3\cdot\mathcal{L}_{div} \tag{9}$$

where λ_{1-3} are the balance weights of the three loss terms.

4. Experiments

4.1. Experimental Settings

Datasets. MARS dataset [52] is one of the existing largest video benchmark, consisting of 1,261 identities and 20,715 video sequences. The training set contains 625 identities and the testing set contains 636 identities. iLIDS-VID dataset [42] is a small-scale benchmark. It consists of 600 video sequences of 300 different identities, each of which has two sequences captured by two non-overlapping cameras. It is randomly split into a training set with 150 identities and a testing set with the remaining 150 identities.

Evaluation Metrics. We adopt the standard metrics, *i.e.*, Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP), to evaluate the performance of different person Re-ID algorithms.

Implementation Details. We randomly sample $T = 6$ frames from a variable-length sequence as a input clip. Each mini-batch has 8 identities and 4 video clips for each identity. We resize all video frames to 256×128 pixels, which are normalized with $1.0/256$. We then apply image-level data augmentation to each video clip, including random horizontal flipping and random erasing probability. ResNet-50 [12] pre-trained on ImageNet is used as the backbone network. The last stride of ResNet-50 is set to 1. The Adam optimizer is adopted with the initial learning rate lr of $3e^{-4}$ and the weight decay of $5e^{-4}$. We train our model for 240 epochs in total. The learning rate lr is decayed by 10 after every 60 epochs. H, W and C are 16, 8 and 2048, respectively. The numbers of 3D-GCLs is set to $L = 2$. α in Eq. 7 is set to 0.3, and λ_{1-3} in Eq. 9 are all set to 1. During inference, $\boldsymbol{V}_f^a$ is used as the final video representation for calculating the similar scores.

4.2. Comparison to State-of-the-Arts

Results on MARS. In Table 1, we compare the proposed method with 13 state-of-the-art methods on MARS dataset. The first two methods belong to the expansion of image-based person Re-ID. We can observe that CTL achieves 91.4% Rank-1 accuracy and 86.7% mAP, surpassing the current state-of-the-art methods by a large margin. It improves the 2nd best method AFA [4] by 1.2% Rank-1 accuracy and MGH [44] by 0.9% mAP, respectively. The comparison clearly demonstrates the effectiveness and superiority of CTL for exploring hierarchical spatial-temporal dependencies and structural information among body parts from videos. Note that, compared with other graph-based methods, including AGRL[43], STGCN [46] and MGH [44], CTL obtains better results in terms of Rank-1 accuracy and mAP. The main reason for the boosting is two aspects: 1) using context-reinforced topology to construct graph instead of pair-wise feature affinity; 2) the advantage of modeling high-order spatial-temporal correlation by 3D-

Table 1. Performance comparison to the state-of-the-art methods on MARS dataset.

Method	Rank-1	Rank-5	Rank-20	mAP
CNN+XQDA [52]	68.3	82.6	89.4	49.3
TriNet [13]	79.8	91.36	-	67.7
STAL[3]	82.2	92.8	98.0	73.5
STAN [21]	82.3	-	-	65.8
COSAM[37]	84.9	95.5	97.9	79.9
VRSTC [15]	88.5	96.5	97.4	82.3
RGSAT [22]	89.4	96.9	98.3	84.0
AGRL [43]	89.8	96.1	97.6	81.1
TCLNet [14]	89.8	-	-	85.1
STGCN [46]	90.0	96.4	98.3	83.7
MGH [44]	90.0	96.7	98.5	85.8
AP3D[11]	90.1	-	-	85.1
AFA [4]	90.2	96.6	-	82.9
CTL	**91.4**	**96.8**	**98.5**	**86.7**

Table 2. Performance comparison to the state-of-the-art methods on iLIDS-VID dataset.

Method	Rank-1	Rank-5	Rank-20
CNN+XQDA [52]	53.0	81.4	95.1
RCNet [29]	58	84.0	96.0
COSAM[37]	79.6	95.3	-
STAN [21]	80.2	-	-
STAL[3]	82.8	95.3	98.8
VRSTC [15]	83.4	95.5	99.5
AGRL [43]	83.7	95.4	99.5
MGH [44]	85.6	97.1	99.5
RGSAT [22]	86.0	**98.0**	99.4
TCLNet [14]	86.6	-	-
AP3D[11]	86.7	-	-
FGRA [6]	88.0	96.7	99.3
AFA [4]	88.5	96.8	99.7
CTL	**89.7**	97.0	**100.0**

GCL and CS-GCL.

Results on iLIDS-VID. Table 2 reports the performance of our approach with 13 state-of-the-art methods on iLIDS-VID dataset. CNN+XQDA [52] and RCNet [29] are the straightforward expansion method of image-based person Re-ID. It can be seen that CTL obtains the best performance of 89.7% Rank-1 accuracy and 100.0% Rank-20 accuracy. It beats AFA [4] on Rank-1 and Rank-20 accuracy by 1.2% and by 0.3%, respectively. The comparison demonstrates the advantage of spatial-temporal feature learning by CTL, and the applicability of CTL for a small-scale dataset.

4.3. Ablation Studies

Effectiveness of Components. Table 3 reports the experimental results of the ablation studies for CTL. Basel denotes using the backbone network with the key-points estimator to learn the global and multi-scale part features. Basel+ContRe denotes using the context-reinforced topology to structure multi-scale frame-wise graphs, and applying original GCN to learn the refined multi-scale part features and global feature. Basel+ContRe+3D refers to CTL using addition operation to replace CS-GCL for learning the fused part feature and the global feature. Basel+ContRe+3D+CS refers to the whole framework of CTL. Compared with Basel, Basel+ContRe boosts Rank-1 accuracy and mAP by 0.8% and 2.7%. This indicts that the effectiveness of the context-reinforced topology to capture the intrinsic relationship among body parts for enhancing feature representation. Moreover, Basel+ContRe+3D improves Basel+ContRe by 1.3% Rank-1 accuracy and 0.6% mAP, which verifies the effectiveness of 3D-GCL for allowing direct cross-spacetime information propagation to enrich part features. By utilizing the cross-scale graph convolutional layer, Basel+ContRe+3D+CS achieves the best performance. This demonstrates that CS-GCL effectively capture diverse visual semantic across multiple scales to integrate them into a comprehensive representation.

Analysis of Context-Reinforced Topology. The results in Table 4 show the influence of different components of the context-reinforced graph topology. CTL-$\boldsymbol{A}_s^p$, CTL-$\boldsymbol{A}_s^p\boldsymbol{A}_s^m$, CTL-$\boldsymbol{A}_s^p\boldsymbol{A}_s^m\boldsymbol{A}_s^c$, denote using $\boldsymbol{A}_s^m$, $\boldsymbol{A}_s^p+\boldsymbol{A}_s^m$ and $\boldsymbol{A}_s^p+\boldsymbol{A}_s^m+\boldsymbol{A}_s^c$ to measure the dependency between two nodes, respectively. By comparing CTL-$\boldsymbol{A}_s^p$ and CTL-$\boldsymbol{A}_s^p\boldsymbol{A}_s^m$, we can conclude that $\boldsymbol{A}_s^m$ improves the flexibility of graph topology and captures more complex spatial-temporal correlation. CTL-$\boldsymbol{A}_s^p\boldsymbol{A}_s^m\boldsymbol{A}_s^c$ achieves remarkable performance improvement as compared to CTL-$\boldsymbol{A}_s^p\boldsymbol{A}_s^m$, which means $\boldsymbol{A}_s^c$ is complementary to physical topology, and mines potential connections that are informative by considering global contextual information of all nodes.

Analysis of 3D-GCL. We conduct the experiments to analyze the influence of the number layer L and window size τ for 3D-GCL. In Figure 7(a), we can observe that the best Rank-1 accuracy of 91.4% and mAP of 86.7% are obtained when $L = 2$. This implies that one-layer 3D-GCL has insufficient capability for capturing complex spatial-temporal information, whilst three-layer 3D-GCLs bring more training parameters, result in hard optimizing and performance degradation. Thus, L is set to 2. In Figure 7(b), $\tau = 3$ obtains superior performance as compared to $\tau = 1$ due to utilizing the temporal complementary information from local temporal neighborhood nodes, but the gain diminishes when $\tau = 5$ as the discriminative clues in aggregated features are counteracted due to the over-sized local temporal neighborhood nodes. Thus, τ is set to 3.

Table 3. Evaluation of the effectiveness of each component of CTL on MARS dataset.

Model	Rank-1	Rank-5	Rank-20	mAP
Basel	88.6	96.1	97.9	82.7
Basel+ContRe	89.4	95.6	98.2	85.4
Basel+ContRe+3D	90.7	96.4	98.4	86.0
Basel+ContRe+3D+CS	91.4	96.8	98.5	86.7

Table 4. Evaluation of the influence of different components of the context-reinforced topology on MARS dataset.

Model	Rank-1	Rank-5	Rank-20	mAP
CTL-$\boldsymbol{A}_s^p$	90.4	96.4	98.5	86.1
CTL-$\boldsymbol{A}_s^p \boldsymbol{A}_s^m$	90.9	96.3	98.6	86.3
CTL-$\boldsymbol{A}_s^p \boldsymbol{A}_s^m \boldsymbol{A}_s^c$	91.4	96.8	98.5	86.7

Table 5. Evaluation of the influence of CS-GCL with different settings on MARS dataset.

Model	Rank-1	Rank-5	Rank-20	mAP
CS-GCL($\boldsymbol{M}=1$)	91.4	96.8	98.5	86.7
CS-GCL($\boldsymbol{M}=2$)	90.1	96.5	98.3	85.5
CS-GCL-$\boldsymbol{s}_3$	90.4	96.4	98.4	85.9
CS-GCL-$\boldsymbol{s}_3\boldsymbol{s}_1$	90.7	96.6	98.4	85.8
CS-GCL-$\boldsymbol{s}_3\boldsymbol{s}_1\boldsymbol{s}_2$	91.4	96.8	98.5	86.7

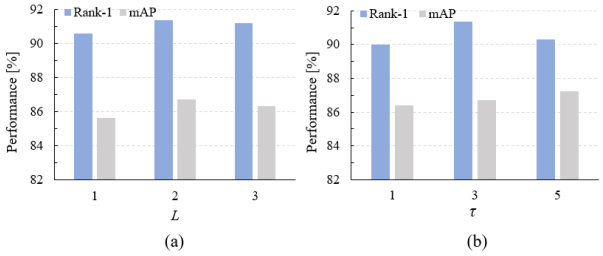

Figure 7. Analysis on the influence of different hyperparameters for 3D-GCL, (a) the number layer L; (b) windows size τ.

Analysis of CS-GCL. In Table 5, we investigate the influence of the number M of CS-GCLs for transferring each granularity-level part features from one scale to another one, and analyze the performance of fusing different scales of part features. We can observe that the two-layer CS-GCLs (denotes using another CS-GCL after the first 3D-GCL) obtain performance degradation over one-layer CS-GCL. It indicates that two-layer CS-GCLs tend to fuse much redundant information, which weaken the representational capability. Moreover, CS-GCL-$\boldsymbol{s}_3\boldsymbol{s}_1\boldsymbol{s}_2$ achieves better results over CS-GCL-$\boldsymbol{s}_3\boldsymbol{s}_1$ and CS-GCL-$\boldsymbol{s}_3$ by combing more granularity-level part features. The improvement verifies CS-GCL can effectively mine the distinct patterns from each scale and enhance feature representation by fusing the complementary information among them.

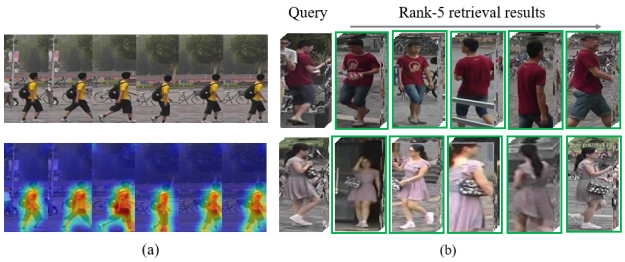

Figure 8. (a) Visualization of the learned feature maps; (b) Visualization of some retrieval results by CTL.

Visualization Results. We visualize the learned feature respond maps of one video sequence by Grad-CAM [32]. In Figure 8(a), we can observe that the feature maps from different video frames of a pedestrian have stronger response on the same discriminative regions, which verifies that CTL can extract aligned discriminative clues by modeling cross-scale spatial-temporal correlation. Figure 8(b) shows the retrieval results of two pedestrians by CTL. We can observe that Rank-5 retrieval results by CTL are all matching. This indicates CTL effectively alleviates the problem of misalignment and occlusion, viewpoint variation, *etc*.

5. Conclusion

In this work, we propose a novel Spatial-Temporal Correlation and Topology Learning framework (CTL) for video-based person re-identification to learn discriminative and robust representation. CTL utilizes a key-points estimator to extract multi-scale part features as graph nodes. A context-reinforced topology is then explored to structure multi-scale graphs by considering global contextual information and physical connection of human body. Moreover, a 3D graph convolution and a cross-scale graph convolution are designed, and performed on the multi-scale graphs. They facilitate direct cross-spacetime and cross-scale information propagation among graph nodes, and model complex relation and structural information to refine pedestrian representation. Extensive experiments on two video datasets validate the effectiveness of the proposed method.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grand 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038 and U19B2023, and China Postdoctoral Science Foundation Funded Project under Grant 2020M671898.

References

[1] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, pages 8649–8658, 2018. 2, 3

[2] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, pages 1268–1277, 2016. 2

[3] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Spatial-temporal attention-aware learning for video-based person re-identification. *IEEE Transactions on Image Processing*, 28(9):4192–4205, 2019. 2, 7

[4] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, pages 660–676, 2020. 6, 7

[5] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *CVPR*, pages 3300–3310, 2020. 2

[6] Zengqun Chen, Zhiheng Zhou, Junchu Huang, Pengyu Zhang, and Bo Li. Frame-guided region-aligned representation for video person re-identification. In *AAAI*, pages 10591–10598, 2020. 2, 7

[7] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, pages 5177–5186, 2019. 3

[8] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, volume 33, pages 8287–8294, 2019. 1, 2

[9] Changxin Gao, Yang Chen, Jin-Gang Yu, and Nong Sang. Pose-guided spatiotemporal alignment for video-based person re-identification. *Information Sciences*, 527:176–190, 2020. 2

[10] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. In *BMVC*, pages 1320–1329, 2018. 1, 2

[11] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, 2020. 1, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 2, 7

[14] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, pages 600–616, 2020. 2, 7

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, pages 7183–7192, 2019. 2, 7

[16] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In *CVPR*, pages 14084–14094, 2020. 2

[17] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, and Wei Zhang. Illumination-invariant person re-identification. In *ACM MM*, pages 365–373, 2019. 2

[18] Xinyang Jiang, Yifei Gong, Xiaowei Guo, Qize Yang, Feiyue Huang, Wei-Shi Zheng, Feng Zheng, and Xing Sun. Rethinking temporal fusion for video-based person re-identification on semantic and time aspect. In *AAAI*, pages 11133–11140, 2020. 1, 2

[19] Michael J Jones and Sai Rambhatla. Body part alignment and temporal attention for video-based person re-identification. In *BMVC*, 2019. 2

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3, 5

[21] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. 2, 7

[22] Xingze Li, Wengang Zhou, Yun Zhou, and Houqiang Li. Relation-guided spatial attention and temporal refinement for video-based person re-identification. In *AAAI*, pages 11434–11441, 2020. 1, 2, 7

[23] Yaoyu Li, Hantao Yao, Lingyu Duan, Hanxing Yao, and Changsheng Xu. Adaptive feature fusion via graph neural network for person re-identification. In *ACM MM*, pages 2115–2123, 2019. 2, 3

[24] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 2

[25] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, pages 7202–7211, 2019. 1

[26] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. Dense 3d-convolutional neural network for person re-identification in videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1s):1–19, 2019. 1

[27] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *ACM MM*, pages 192–196. ACM, 2016. 2

[28] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. 3

[29] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. 1, 2, 7

[30] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019. 2

[31] Deqiang Ouyang, Yonghui Zhang, and Jie Shao. Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks. *Pattern Recognition Letters*, 117:153–160, 2019. 2

[32] 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 *ICCV*, pages 618–626, 2017. 8

[33] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, pages 486–504, 2018. 1, 2, 3

[34] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *CVPR*, pages 7912–7921, 2019. 4

[35] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *CVPR*, pages 1711–1719, 2020. 3

[36] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, October 2019. 1, 2

[37] Arulkumar Subramaniam, Athira Nambiar, Anurag Mittal, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, pages 562–572, 2019. 2, 7

[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 3

[39] 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 *ECCV*, pages 480–496, 2018. 1, 2, 3

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 6

[41] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, pages 7134–7143, 2019. 1

[42] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. Springer, 2014. 2, 6

[43] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing*, 29:8821–8830, 2020. 1, 2, 3, 6, 7

[44] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *CVPR*, pages 2899–2908, 2020. 2, 3, 6, 7

[45] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *CVPR*, pages 2158–2167, 2019. 2, 3

[46] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, pages 3289–3299, 2020. 2, 3, 6, 7

[47] Zheng-Jun Zha, Jiawei Liu, Di Chen, and Feng Wu. Adversarial attribute-text embedding for person search with natural language query. *IEEE Transactions on Multimedia*, 22(7):1836–1846, 2020. 1

[48] Wei Zhang, Shengnan Hu, Kan Liu, and Zhengjun Zha. Learning compact appearance representation for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2442–2452, 2018. 2

[49] Xikun Zhang, Chang Xu, and Dacheng Tao. Context aware graph convolution for skeleton-based action recognition. In *CVPR*, pages 14333–14342, 2020. 3

[50] Yiheng Zhang, Dong Liu, and Zheng-Jun Zha. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In *ICME*, pages 1386–1391, 2017. 2

[51] 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 *CVPR*, pages 1077–1085, 2017. 2

[52] 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 *ECCV*, pages 868–884. Springer, 2016. 1, 2, 6, 7

[53] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *ECCV*, 2020. 2