Adaptive Graph Representation Learning for Video Person Re-identification

Yiming Wu, Omar El Farouk Bourahla, Xi Li*, Fei Wu, and Qi Tian, *Fellow, IEEE*

***Abstract*—Recent years have witnessed a great development of deep learning based video person re-identification (Re-ID). A key factor for video person Re-ID is how to effectively construct discriminative video feature representations for the robustness to many complicated situations like occlusions. Recent part-based approaches employ spatial and temporal attention to extract the representative local features. While the correlations between the parts are ignored in the previous methods, to leverage the relations of different parts, we propose an innovative adaptive graph representation learning scheme for video person Re-ID, which enables the contextual interactions between the relevant regional features. Specifically, we exploit pose alignment connection and feature affinity connection to construct an adaptive structure-aware adjacency graph, which models the intrinsic relations between graph nodes. We perform feature propagation on the adjacency graph to refine the original regional features iteratively, the neighbor nodes information is taken into account for part feature representation. To learn the compact and discriminative representations, we further propose a novel temporal resolution-aware regularization, which enforces the consistency among different temporal resolutions for the same identities. We conduct extensive evaluations on four benchmarks, i.e. iLIDS-VID, PRID2011, MARS, and DukeMTMC-VideoReID, the experimental results achieve the competitive performance which demonstrates the effectiveness of our proposed method.**

***Index Terms*—Video Person Re-Identification, Graph Neural Network, Consistency**

I. Introduction

As an important and challenging problem in computer vision, *person re-identification* (Re-ID), aims at precisely retrieving the same identities from the gallery with a person of interest as a query given, has a wide range of applications in intelligent surveillance and video analysis [1]. Typically, person Re-ID is carried out in the domain of individual images without capturing the temporal coherence information. More recently, a number of video person Re-ID approaches [2]–[18] emerge to directly perform the person context modeling at the video level, which is more fit for practical use with more visual cues exploited for coping with complicated circumstances.

In the literature, most existing methods for video person Re-ID first extract the feature vector frame by frame and generate the video-level feature representation by temporal aggregation, and subsequently compare them in a particular metric space. Although recent deep learning based methods have made notable progress, it remains challenging due to occlusion, viewpoints, illumination, and pose variation in the video. To address these issues, recent studies [4]–[6], [19], [20] concentrate on aggregating the effective images' regions with attention mechanisms. However, under the circumstances of complicated situations (e.g. occlusion and pose variations), these approaches are often incapable of effectively utilizing the intrinsic context relations between person parts across frames, which play an important role in learning robust video representations. For instance, if the body parts are occluded in the first frame, the appearance cues and contextual information

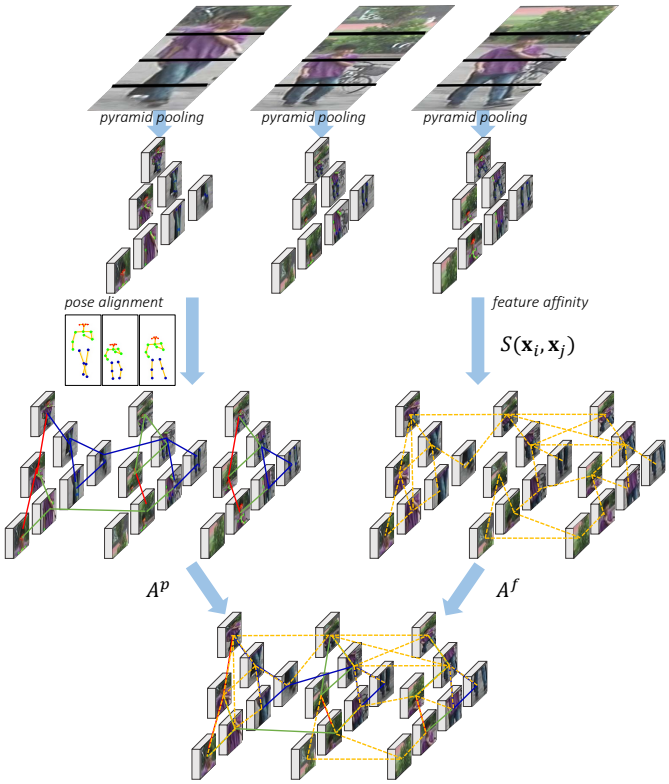

Fig. 1. Overview of graph construction in our proposed method. 1) The feature map from each frame is processed by a pyramid pooling module, the extracted regional features are treated as the graph nodes. 2) The pose alignment adjacency graph A^p (colorful solid line) is constructed by connecting the regions containing the same human part. 3) The feature affinity adjacency graph A^f (yellow dotted line) is constructed by measuring the affinity of regional features. 4) The adaptive structure-aware adjacency graph is built by combining two graphs. **Best viewed in color, some graph edges are omitted for clarity.**

Y. Wu, Omar, F. Wu are with College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: ymw, obourahla@zju.edu.cn wufei@cs.zju.edu.cn).

X. Li*(corresponding author) is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China, and also with the Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou 310027, China (e-mail: xilizju@zju.edu.cn).

Q. Tian is with the Department of Computer Science, University of Texas, San Antonio, TX 78249-1604 USA (e-mail: qitian@cs.utsa.edu)

from the other frames are complementary. Hence, how to adaptively perform relation modeling and contextual information propagation among spatial regions is a key issue to solve in video person Re-ID.

Motivated by the aforementioned observations, we propose an adaptive graph learning scheme to model the contextual relations and propagate complementary information simultaneously. As shown in Figure I, we construct two kinds of relations named pose alignment connection and feature affinity connection between the spatiotemporal regions. Specifically, 1) pose alignment connection: the regions containing the same part are connected to align the spatial regions across the frames. With the pose alignment connection, we are capable of capturing the relations between human body parts; and 2) feature affinity connection: we define the edges by measuring the visual correlations of extracted regional features. With the feature affinity connection, we can model the visually semantic relationships between regional features accurately.

By combining these two complementary relation connections, we obtain an adaptive structure-aware adjacency graph. Then we capture the contextual interactions on the graph structure via graph neural network (GNN). As a result, the effective and discriminative messages aggregated from neighbors are used to refine the original regional features. With feature propagation, the discriminative power of the informative regional features is enhanced while the noisy parts are weakened.

Based on the observations in [21], the visual cues in the video are rich yet possibly redundant, and some keyframes are sufficient to represent the long-range video. As a consequence, we propose a novel regularization, which enforces the consistency among different temporal resolutions, to learn the temporal resolution invariant representation. Specifically, the frame-level features are randomly selected as the subsequences, and input into the attention module, then the output video representations are enforced to be close to each other in the metric space.

Overall, the main contributions of this work are summarized as follows:

- We propose an adaptive structure-aware spatiotemporal graph representation based on two types of graph connections for relation modeling: pose alignment connection and feature affinity connection. By combining these two relation connections, the adaptive structure-aware graph representation is capable of well capturing the semantic relations between regions across frames.
- We propose a novel regularization to learn the temporal resolution invariant representation, which is compact and captures the discriminative information in the sequence.

We conduct extensive experiments on four widely used benchmarks (i.e. iLIDS-VID, PRID2011, MARS, and DukeMTMC-VideoReID), and the experimental results demonstrate the effectiveness of our proposed method.

II. Related Work

Person Re-ID. Person Re-ID in still images is widely explored [22]–[32]. Currently, the researchers start to focus on video-based person Re-ID [2], [33]. Facilitated by deep learning technique, impressive progress has been observed with video person Re-ID recently. McLaughlin *et al.* [2] and Yan *et al.* [33] employ RNN to model the inter-sequence dependency and aggregate the features extracted from the video frames with average pooling or max pooling. Zhou *et al.* [3] separately model the spatial and temporal coherence with two RNN networks: the temporal model (TAM) focuses on discriminative frames and spatial model (SRM) integrates the contexture at different locations for better similarity evaluation. Wu *et al.* [20] extend GRU with attention mechanism to selectively propagate relevant features and memorize their spatial dependencies through the network. Dai *et al.* [9] propose a S^2TN network to address the pose alignment and combine the bi-directional LSTM with residual learning to perform temporal residual learning.

Recently, the attention networks are widely studied for temporal feature fusion. In [4], [5], the discriminative frames are selected with attention temporal pooling, where each frame is assigned with a quality score and then fused to a final video representation. Similarly, Zhang *et al.* [21] employ reinforcement learning to train an agent to verify whether the pair of images are same or different, and the Q value is a good indicator of the difficulty of image pairs. In [6], [13], [15], [19], the authors extend the temporal attention to spatiotemporal attention to select informative regions and achieve the impressive improvements. Chen *et al.* [12] leverage the body joints to attend to the saliency parts of the person in the video to extract the discriminative local features in a siamese network. And in [7], [11], the video representation is generated by considering not only intra-sequence influence but also inter-sequence mutual information. Different from the previous 2D CNN based methods, 3D convolution neural network (3D CNN) is also adopted to address the video person Re-ID [10], [14], [16]. Wu *et al.* [10] adopt 3D CNN and 3D pooling to aggregate the spatial and temporal cues simultaneously. Li *et al.* [14] propose a variant of ResNet by inserting multi-scale 3D (M3D) layer and residual attention layer (RAL) into the ResNet. Similarly, Liu *et al.* [16] incorporate non-local modules with ResNet50 as the Non-local Video Attention Network (NVAN), and propose a spatially and temporally efficient variant. Moreover, attributions are utilized to generate the confidence as the weight for sub-features extracted from video frames in [34]. The generative models are adopted to address the occlusion and pose variant in [17], [18].

Graph Models. Graph models are utilized in several computer vision tasks, and Graph Neural Networks (GNN) is introduced in [35] to model the relations between graph nodes, and a large number of the variants [36]–[41] are proposed. In recent, Re-ID methods [42]–[46] combined with graph models are also explored. Cheng *et al.* [42] formulate the structured distance relationships into the graph Laplacian form to take advantages of the relationships among training samples. In [43], an algorithm that maps the ranking process to a problem in graph theory is proposed. Shen *et al.* [44] leverage the similarities between different probe-gallery pairs for updating the features extracted from images. Chen *et al.* [45] involves multiple images to model the relationships among the local and global similarities in a unified CRF. Yan

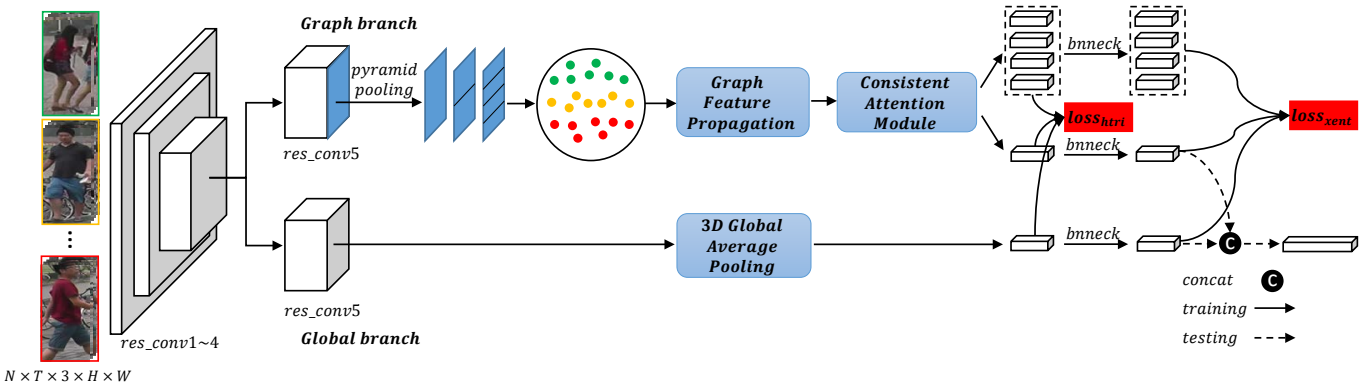

Fig. 2. The overall architecture of our proposed method. 1) T frames are sampled from a long-range video with a restricted random sampling method. 2) In graph branch, for the output of each image, pyramid pooling is used to extract the $N \times d$-dimension feature, where N represents the number of regions, the feature vector for each region has d dimensions. 3) The extracted feature vectors are treated as the graph nodes, we then employ GNN to perform feature propagation on the graph iteratively in graph feature propagation module. 4) We carry out the attention module to generate the discriminative video representation, the subsequences are randomly selected and forward into the attention module to learn a consistent video representation. 5) Feature vectors from graph branch and global branch are concatenated for testing.

et al. [46] formulate the person search as a graph matching problem, and solve it by considering the context information in the probe-gallery pairs. To address the unsupervised Re-ID problem, Ye *et al.* [47] involves the graph matching into an iteratively updating procedure for a robust label estimation.

In a nutshell, the graph model based methods in Re-ID usually build up a graph to represent the relationships among training samples, where the graph nodes are images or videos. While in our proposed method, the graph is dynamically learned with prior knowledge to model the intrinsic contextual relationships among the regions in an image sequence, the local, global and structure information are propagated among the different regional features to learn the discriminative video feature representation.

III. The Proposed Method

A. Overview

Video person Re-ID aims to retrieve the identities from the gallery with the given queries. The overall architecture of our proposed method is illustrated in Figure 2. Given a long-range video for the specific identity, T frames are randomly sampled with restricted sampling method [6], [48], and then grouped as an image sequence $\{I_t\}_{t=1,...,T}$. We first feed them into the ResNet50-based [49] feature extractor module, in which the stride of the first residual block in $conv5$ is set to 1. In the global branch, 3D global average pooling is used for the feature maps and produces a video representation $\mathbf{x}_{gap} \in \mathbb{R}^d$. In the graph branch, we obtain regional features $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{T \cdot N}$ with pyramid pooling [24], where the feature maps are vertically partitioned into 1, 2, and 4 regions[1] in our experiments, and $N = 7$ is the number of regions for an individual frame. Then, we utilize the pose information and feature affinity to construct an adaptive structure-aware adjacency graph, which captures the intrinsic relations between these regional features. In the graph feature propagation module, the regional features are updated iteratively by aggregating the contextual information from neighbors on the graph. Next, we utilize the attention module to yield the video representations. The network is supervised by identification loss and triplet ranking loss together. We will discuss these modules in the following sections.

[1]In this paper, regions and nodes are interchangeably for the same meaning.

B. Adaptive Graph Feature Propagation

As discussed in Section I, the relations between human parts are beneficial for mitigating the impact of complex situations such as occlusion and clutter background. So, how to describe the relationships between different human parts and propagate contextual messages is critical for learning the discriminative video representations. The graph is commonly used to model this kind of relations, and we adopt GNN to leverage the information from the neighborhood.

Adaptive Structure-Aware Adjacency Graph. To depict the relations of human parts, we employ the pose information and feature affinity to construct an adaptive structure-aware adjacency graph $G = \{V, A\}$. $V = \{v_i\}_{i=1}^{T \cdot N}$ is the vertex set containing $T \cdot N$ nodes, where each node v_i corresponds to a spatial region in the frame. To define the edge $A \in \mathbb{R}^{(T \cdot N) \times (T \cdot N)}$ on the graph, we introduce two types of relations: pose alignment connection and feature affinity connection.

The pose alignment connection is defined by leveraging the human body joints: two regions (nodes) are connected if they contain the same human parts. Formally, we define a set S_i for each region v_i where $S_i \subseteq \{head, body, leg\}$. The pose alignment adjacency graph A^p for the two nodes v_i and v_j is then calculated as follows:

$$A^p_{ij} = \begin{cases} 1 & i \neq j \quad \text{and} \quad |S_i \cap S_j| \neq 0, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where $|\cdot|$ means the cardinality of a set. We obtain S_i with the following procedure, first, we locate the joints of

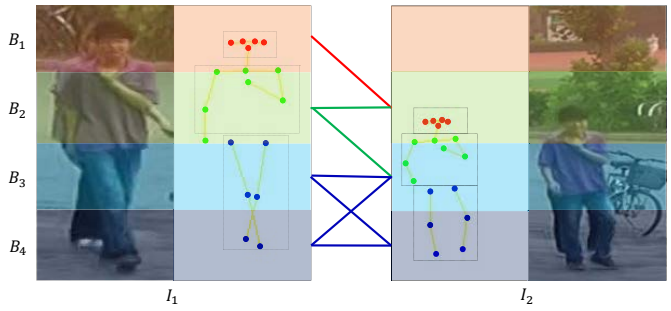

Fig. 3. Construction of pose alignment adjacency graph for spatial regions. The estimated keypoints are separated into three parts: head part (consist of nose, neck, eyes, and ears), body part (consist of shoulders, bows, and wrists), and leg part (consist of hips, knees, and ankles). The spatial regions (nodes) containing the same part are connected in the pose alignment adjacency graph.

the human body by making use of a human pose estimation algorithm, this is illustrated in Figure 3. Then we separate these estimated keypoints into three parts: head part (consist of nose, neck, eyes, and ears), body part (consist of shoulders, bows, and wrists), and leg part (consist of hips, knees, and ankles). For the i-th spatial region, S_i is constitutive of the parts located in this spatial region. We present an example in Figure 3, the feature maps are vertically partitioned into 4 regions B_1, B_2, B_3, B_4. In image I_1 and I_2, the head part is in B_1 and B_2 respectively, so the pose alignment connection between these two nodes is set to 1. Then, we can create the pose alignment adjacency graph A^p.

Pose alignment adjacency connection reflects only the coarse relations between different spatial regions and the recent method [50] shows the dynamic graph could learn better graph representations compared to the fixed graph structure. To describe the fine relations between the regions, we propose to learn an adaptive feature affinity adjacency graph A^f, which aims to capture the affinity between the regions. For two nodes v_i and v_j, the node features are $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively, then the entry of adjacency graph A^f is formulated as follows:

$$\begin{aligned} A^f_{ij} =& S(\mathbf{x}_i, \mathbf{x}_j) \\ =& \frac{2}{e^{\|\mathbf{x}_i - \mathbf{x}_j\|_2} + 1}. \end{aligned} \tag{2}$$

We calculate the edge weight matrix A by combining the pose alignment adjacency matrix and feature affinity matrix:

$$A_{ij} = \frac{1}{1+\gamma}\left(\frac{A^p_{ij}}{\sum_j A^p_{ij}} + \gamma \frac{A^f_{ij}}{\sum_j A^f_{ij}}\right), \tag{3}$$

where γ is the weight parameter to balance the pose alignment adjacency matrix and the feature affinity matrix, and γ is set to 1 in our all experiments.

Graph Feature Propagation Module. After obtaining the graph, we perform contextual message propagation to update original spatial regional features iteratively.

As shown in Figure 4, we employ Graph Neural Network (GNN) [35] to aggregate the information from neighbors for each node. In the graph feature propagation module, we stack

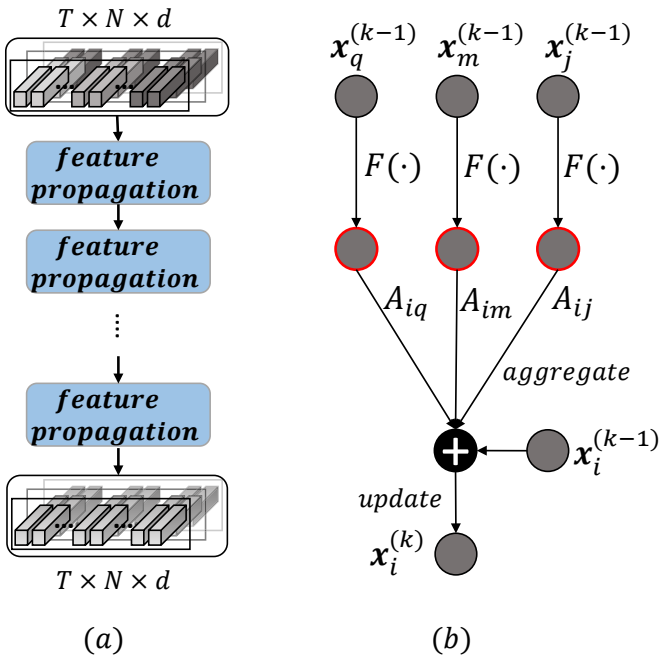

Fig. 4. (a) Graph feature propagation module. Given the adaptive pose alignment adjacency graph A, the original spatial regional features are updated iteratively through the feature propagation layers. (b) Feature propagation layer. The features from neighbors are processed by an fc layer $F(\cdot)$, and then aggregated by the weights from adjacency graph. The operation is defined in Equation 4.

L graph feature propagation layers, in the l-th layer, the aggregation and updating operations are defined as follows:

$$\mathbf{x}_i^{(l)} = (1-\alpha)\mathbf{x}_i^{(l-1)} + \alpha \sum_{j=1}^{T \cdot N} A_{ij}^{(l)} F^{(l)}(\mathbf{x}_j^{(l-1)}) \tag{4}$$

where $i \in \{1, 2, \ldots, T \cdot N\}$, $l \in \{1, 2, \ldots, L\}$, and $\mathbf{x}_i^{(l)}$ stands for the refined regional feature output from l-th feature propagation layer and $\mathbf{x}_i^{(0)} = \mathbf{x}_i$ is the original node feature, $F^{(l)}(\cdot)$ is the combination of an FC-layer and batch normalization layer to encode the contextual messages from neighbors, $A^{(l)}$ refers to the adaptive structure-aware adjacency graph, and α is used to balance the aggregated feature and original feature, which is set as 0.1 in our experiments. The output from graph feature propagation module is denoted as $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_{T \cdot N}]$, where $\hat{\mathbf{x}}_i \in \mathbb{R}^d$ is the updated regional feature vector.

Temporal Attention Module Given the updated regional features $\hat{\mathbf{X}}$, we perform a simple yet effective spatio-temporal attention [15] to obtain the video representation, which is calculated as:

$$\mathbf{x}_{graph} = \sum_{i=1}^{T \cdot N} \frac{\|\hat{\mathbf{x}}_i\|_1}{\sum_j \|\hat{\mathbf{x}}_j\|_1} \hat{\mathbf{x}}_i. \tag{5}$$

As discussed in Section I, in order to model the consistency of subsequences, we randomly select $T - i$ frames from the image sequence, where $i = 1, 2, \ldots, T_s$. And then feed the feature vectors of these subsequence frames into the temporal attention module to obtain video representations

$\mathbf{x}_{graph,1}, \ldots, \mathbf{x}_{graph,T_s}$. To keep the consistency of different subsequences, we enforce these video representations to be close to each other in the metric space.

C. Loss Functions

We employ two kinds of losses to jointly supervise the training of parameters: cross entropy loss and soft hard triplet loss [51], the losses are formulated as follows:

$$L_{xent}(\mathbf{x}) = -\frac{1}{P \cdot K} \sum_{i=1}^{P} \sum_{a=1}^{K} log[\frac{e^{\mathbf{W}_{y_{i,a}}{}^{\top} \mathbf{x}_{i,a}}}{\sum_{c=1}^{P \cdot K} e^{\mathbf{W}_c{}^{\top} \mathbf{x}_{i,a}}}], \quad (6)$$

$$L_{htri}(\mathbf{x}) = \sum_{i=1}^{P} \sum_{a=1}^{K} ln(1+exp(\overbrace{\max_{p=1,\cdots,K} D(\mathbf{x}_{i,a}, \mathbf{x}_{i,p})}^{hardest\ positive} - \underbrace{\min_{\substack{n=1,\ldots,K \\ j=1,\ldots,P \\ j \neq i}} D(\mathbf{x}_{i,a}, \mathbf{x}_{j,n})}_{hardest\ negtive})), \quad (7)$$

where P and K are respectively the number of identities and sampled images of each identity. So there are $P \cdot K$ images in a mini-batch, $\mathbf{x}_{i,a}$, $\mathbf{x}_{i,p}$ and $\mathbf{x}_{j,n}$ are the features extracted from the anchor, positive and negative samples respectively, $D(\cdot)$ is the L2-norm distance for two feature vectors.

For the output from global branch, we have two losses l_{xent}^{global} and l_{htri}^{global}:

$$l_{xent}^{global} = L_{xent}([BN(\mathbf{x}_{gap}^{(1)}), \ldots, BN(\mathbf{x}_{gap}^{(P \cdot K)})]), \quad (8)$$

$$l_{htri}^{global} = L_{htri}([\mathbf{x}_{gap}^{(1)}, \ldots, \mathbf{x}_{gap}^{(P \cdot K)}]), \quad (9)$$

for the output from graph branch, we have two losses l_{xent}^{graph} and l_{htri}^{graph} similarily:

$$l_{xent}^{graph} = L_{xent}([BN(\mathbf{x}_{graph}^{(1)}), \ldots, BN(\mathbf{x}_{graph,T_s}^{(1)}), \ldots, BN(\mathbf{x}_{graph}^{(P \cdot K)}), \ldots, BN(\mathbf{x}_{graph,T_s}^{(P \cdot K)})]), \quad (10)$$

$$l_{htri}^{graph} = L_{htri}([\mathbf{x}_{graph}^{(1)}, \mathbf{x}_{graph,1}^{(1)}, \ldots, \mathbf{x}_{graph,T_s}^{(1)}, \ldots, \mathbf{x}_{graph}^{(P \cdot K)}, \mathbf{x}_{graph,1}^{(P \cdot K)}, \ldots, \mathbf{x}_{graph,T_s}^{(P \cdot K)}]), \quad (11)$$

where $BN(\cdot)$ is the BNNeck introduced in [52], $[\cdot]$ means concatenation. The total loss is the summation of the four losses:

$$l_{total} = l_{xent}^{global} + l_{htri}^{global} + l_{xent}^{graph} + l_{htri}^{graph} \quad (12)$$

IV. Experiments

A. Datasets

PRID2011 [53] dataset consists of person videos from two camera views, containing 385 and 749 identities, respectively. Only the first 200 people appear in both cameras. The length of image sequence varies from 5 to 675 frames, but we use only the sequences whose frame number is larger than 21.

iLIDS-VID [54] dataset consists of 600 image sequences of 300 persons. For each person, there are two videos with the sequence length ranging from 23 to 192 frames with an average duration of 73 frames.

MARS dataset [55] is the largest video-based person re-identification benchmark with 1,261 identities and around 20,000 video sequences generated by DPM detector and GMMCP tracker. The dataset is captured by six cameras, each identity is captured by at least 2 cameras and has 13.2 sequences on average. There are 3,248 distracter sequences in the dataset, it increases the difficulty of Re-ID.

DukeMTMC-VideoReID dataset is another large scale benchmark dataset for video-based person Re-ID, which is derived from the DukeMTMC dataset [56] and re-organized by Wu *et al.* [57]. The DukeMTMC-VideoReID dataset contains totally 4,832 tracklets and 1,812 identities, it is separated into 702, 702 and 408 identities for training, testing and distraction. In total, it has 369,656 frames of 2,196 tracklets for training, and 445,764 frames of 2,636 tracklets for testing and distraction. Each tracklet has 168 frames on average, and the bounding boxes are annotated manually.

B. Evaluation Metrics

For evaluation, we employ the standard metrics used in person Re-ID literature: cumulative matching characteristic (CMC) curve and mean average precision (mAP). CMC curve judges the ranking capabilities of the Re-ID model, mAP reflects the true ranking results while multiple ground-truth sequences exist. For PRID2011 and iLIDS-VID datasets, we follow the evaluation protocol used in [53]. Each dataset is divided into two parts for training and testing, the final accuracy is the average of "10-fold cross validation", only CMC accuracy is reported in PRID2011 and iILIDS-VID because of the equivalence of CMC and mAP on these two datasets. For MARS and DukeMTMC-VideoReID dataset, both CMC and mAP are reported.

C. Implementation Details

Settings. Our experiments are implemented with Pytorch and four TITAN X GPUs. ResNet50 [49] is first pre-trained on ImageNet, and the input images are all resized to 256×128. In the training stage, we employ restricted random sampling strategy [6] to randomly sample $T = 8$ frames from every video and group them into a tracklet. We update the parameters by employing ADAM [58] with a learning rate of 1×10^{-4} and weight decay of 5×10^{-4}. We train the network for 300 epochs, the learning rate decays to $\frac{1}{10}$ every 100 epochs. For batch hard triplet loss, we set $P = 4$ and $K = 4$ in our experiments. In the temporal attention module, T_s is set as 3. In the testing stage, cosine distance between the representations is calculated for ranking, a video containing T_v frames is split into T chunks firstly, then we make use of two kinds of strategies: 1) the first image is collected as an image sequence to represent this video; 2) In each chunk, i-th frames are grouped as an image sequence, we can obtain $\lceil \frac{T_v}{T} \rceil$ image sequences, and the video representations are averaged as a single video representation. In our experiments, the first strategy is fast and the second strategy is more accurate.

Pose Estimation. To generate the human body joints, we adopt AlphaPose [59] as our pose estimation algorithm, which achieves state-of-the-art performance in human pose

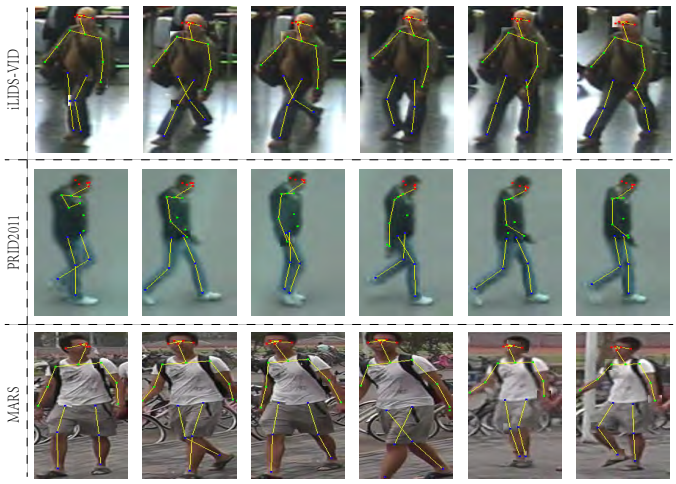

Fig. 5. Keypoint detection examples. The human body joints are marked by colorful dots, and linked by yellow lines. **Best viewed in color and zoom in.**

estimation. There are total of 18 joints detected for each person image, which costs 20ms on average. And we show some examples in Figure 5.

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

To validate the effectiveness of our proposed method, we compare our proposed method with several state-of-the-art methods on PRID2011, iLIDS-VID, MARS, and DukeMTMC-VideoReID, include RCN [2], IDE+XQDA [55], RFA-Net [33], SeeForest [3], QAN [5], AMOC+EF [60], ASTPN [4], Snippet [7], STAN [6], EUG [57], SDM [21], RQEN [19], PersonVLAD [10], M3D [14], STMP [13], STA [15], TRL+XQDA [9], SCAN [11], STAL [12], and STE-NVAN [16].

Results on MARS and DukeMTMC-VideoReID From Table I and Table II, it is not difficult to find that our proposed method outperforms the existing approaches. On MARS dataset, our proposed method surpasses the previous best approach STE-NVAN [16] by 0.6% and 0.7% in terms of Rank-1 and mAP. And on DukeMTMC-VideoReID, our method achieves the best performance with 97.0% and 95.4% at Rank-1 and mAP accuracy respectively. The results outperform the state-of-the-art STA [15] by a large margin. These experimental results confirm the effectiveness and superiority of our proposed method.

Results on iLIDS-VID and PRID2011 As shown in Table III, results demonstrate the advantages of the proposed method over existing state-of-the-art approaches on iLIDS-VID and PRID2011. Specifically, our proposed method achieves the Rank-1 accuracy of 84.5% and 94.6% on these two datasets, and surpass all the previous approaches without incorporating optical flow. On these two small-scale datasets, by comparing Snnipet and Snippet+OF, we can observe that the motion information provides more reliable features than appearance cues. While even compared to those methods utilizing optical flow, the results of our proposed method is also compatitive.

Method	MARS			
	R1	R5	R20	mAP
IDE+XQDA [55]	65.3	82	89	47.6
SeeForest [3]	70.6	90	97.6	50.7
ASTPN [4]	44	70	81	-
RQEN [19]	73.74	84.9	91.62	71.14
Snippet [7]	81.2	92.1	-	69.4
STAN [6]	82.3	-	-	65.8
SDM [21]	71.2	85.7	94.3	-
EUG [57]	62.67	74.94	82.57	42.45
PersonVLAD [10]	80.8	94.5	**99**	63.4
DSAN+KISSME [20]	73.5	85	97.5	
TRL+XQDA [9]	80.5	91.8	96	69.1
M3D [14]	84.39	93.84	97.74	74.06
STA [15]	86.3	95.7	98.1	80.8
STMP [13]	84.4	93.2	96.3	72.7
SCAN [11]	86.6	94.8	97.1	76.7
STAL [12]	82.2	92.8	98	73.5
STE-NVAN [16]	88.9	-	-	81.2
AMOC+EF† [60]	68.3	81.4	90.6	52.9
Snippet+OF† [7]	86.3	94.7	-	76.1
SCAN+OF† [11]	87.2	95.2	98.1	77.2
Ours	**89.8**	96.1	97.6	81.1
+Test Strategy 2	89.5	**96.6**	97.8	**81.9**

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON MARS DATASET, RANK-1, -5, -20 ACCURACIES(%) AND MAP ARE REPORTED. † REFERS TO OPTICAL FLOW, AND *Test Strategy 2* IS THE SECOND STRATEGY INTRODUCED IN SECTION IV-C.

Method	DukeMTMC-VideoReID			
	R1	R5	R20	mAP
STA [15]	96.2	**99.3**	-	94.9
STE-NVAN [16]	95.2	-	-	93.5
EUG [57]	83.6	94.6	97.6	78.3
VRSTC [18]	95	99.1	-	93.5
Ours	96.7	99.2	99.7	94.2
+Test Strategy 2	**97.0**	**99.3**	**99.9**	**95.4**

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON DUKEMTMC-VIDEOREID DATASET, RANK-1, -5, -20 ACCURACIES(%) AND MAP ARE REPORTED. † REFERS TO OPTICAL FLOW, AND *Test Strategy 2* IS THE SECOND STRATEGY INTRODUCED IN SECTION IV-C.

E. Ablation Study

To analyze the effectiveness of components in our proposed method, we conduct several experiments on MARS dataset. The experimental results are summarized in Table V and Table IV.

Analysis on feature propagation module. We carry out experiments to investigate the effect of varying the number of feature propagation layers. We evaluate the results of stacking 1, 2, and 3 propagation layers based on the model combined with adaptive pose alignment adjacency graph and consistent loss, i.e. $+A^p + A^f + consistent$ in Table V. As shown in Table IV, we find out that the performance is consistent in these settings, where Rank-1 accuracies are all above 89.0% and mAP accuracies are all above 80.5%. In addition, the performance with $K = 2$ surpasses the other settings, so we stack 2 propagation layers in our experiments.

Analysis on components. In Table V, *Baseline* contains only the ResNet backbone and 3D global average pooling, and is supervised by l_{xent}^{gap} and l_{htri}^{gap}, the Rank-1 and mAP accuracy of baseline approach is 87.8% and 78.0% respectively. *+Attention* refers to adopt temporal attention module in

Method	iLIDS-VID			PRID2011		
	R1	R5	R20	R1	R5	R20
IDE+XQDA [55]	53	81.4	95.1	77.3	93.5	99.3
RFA-Net [33]	58.2	85.8	97.9	49.3	76.8	90
RCN [2]	58	84	96	70	90	97
SeeForest [3]	55.2	86.5	97	79.4	94.4	99.3
QAN [5]	68	86.8	97.4	90.3	98.2	**100**
ASTPN [4]	62	86	98	77	95	99
RQEN [19]	76.1	92.9	99.3	92.4	98.8	**100**
Snippet [7]	79.8	91.8	-	88.6	99.1	-
STAN [6]	80.2	-	-	93.2	-	-
SDM [21]	60.2	84.7	95.2	85.2	97.1	99.6
DSAN+KISSME [20]	61.9	86.8	98.6	77	96.4	99.4
TRL+XQDA [9]	57.7	81.7	94.1	87.8	97.4	99.3
M3D [14]	74	94.33	-	94.4	**100**	-
STMP [13]	84.3	96.8	**99.5**	92.7	98.9	99.8
PersonVLAD [10]	69.4	87.6	99.2	87.6	96.1	99.8
SCAN [11]	81.3	93.3	98	92	98	**100**
STAL [12]	82.8	95.3	98.8	92.7	98.8	**100**
VRSTC [18]	83.4	95.5	99.5	-	-	-
AMOC+EF† [60]	68.7	94.3	99.3	83.7	98.3	100
Snippet+OF † [7]	85.4	96.7	-	93	99.3	-
SCAN+OF † [11]	88	96.7	100	95.3	99	100
Ours	83.7	95.4	99.5	93.1	98.7	99.8
+Test Strategy 2	**84.5**	96.7	**99.5**	**94.6**	99.1	**100**

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON PRID2011 AND ILIDS-VID DATASETS, RANK-1, -5, -20 ACCURACIES(%) ARE REPORTED. † REFERS TO OPTICAL FLOW, AND *Test Strategy 2* IS THE SECOND STRATEGY INTRODUCED IN SECTION IV-C. THE APPROACHES UTILIZING OPTICAL FLOW ARE NOT DIRECTLY COMPARED. THE EXPERIMENTAL RESULTS INDICATE THAT OUR PROPOSED METHOD ACHIEVES THE STATE-OF-THE-ART PERFORMANCE.

K	R1	R5	R10	R20	mAP
1	89.3	96.0	97.1	**97.6**	80.8
2	**89.8**	**96.1**	97.0	**97.6**	**81.1**
3	89.5	96.1	**97.1**	97.0	**81.1**

TABLE IV
ANALYSIS ON FEATURE PROPAGATION MODULE. K IS THE NUMBER OF FEATURE PROPAGATION LAYERS. WE USE THE MODEL TRAINED WITH $K = 2$ IN OUR EXPERIMENTS.

Model	R1	R5	R10	R20	mAP
Baseline	87.8	95.3	96.3	97.3	78.0
+Attention	88.0	95.3	96.9	**97.9**	79.4
+A^p	88.7	95.9	96.9	97.8	79.8
+A^p+A^f	89.3	96.1	96.9	97.8	80.4
+A^p+A^f+consistent	**89.8**	**96.1**	**97.0**	97.6	**81.1**

TABLE V
ABLATION STUDY ON MARS DATASET, WE PRESENT RANK-1, -5, -10, -20 ACCURACY(%) AND MAP(%). A^p, A^p+A^f, AND CONSISTENT REPRESENT POSE ALIGNMENT ADJACENCY GRAPH, COMBINED ADJACENCY GRAPH, AND CONSISTENT LOSS RESPECTIVELY. BASELINE CONSISTS OF FEATURE EXTRACTOR AND TEMPORAL AVERAGE POOLING, THE NUMBER OF FEATURE PROPAGATION LAYER IS SET TO 2.

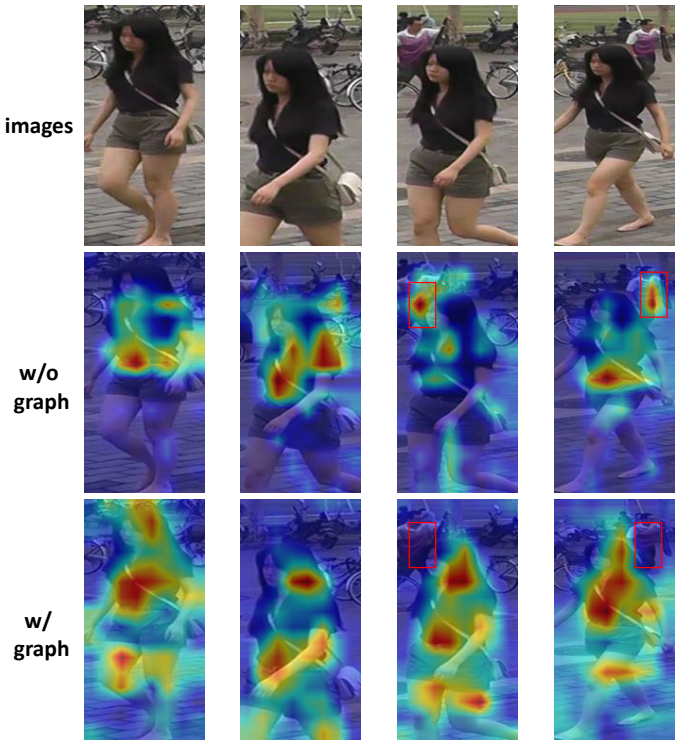

Fig. 6. Qualitative results for the baseline model (w/o graph) and our proposed method (w/ graph). We employ Grad-CAM to visualize the regions focused by the Re-ID model. The first row is the image sequences sampled from videos. In the second and third row, the class activation maps for the baseline model and our proposed method are provided. Compared to the model without graph learning, we find out that our proposed method is robust to occlusion and clutter background. The results for different sequences are separated by the line.

the graph branch, the corresponding performance is 88.0% in Rank-1 and 79.4% in mAP. A^p refers to only adopting the pose alignment adjacency graph in feature propagation module, we stack 2 feature propagation layers in our experiments. Compared to *+Attention*, $+A^p$ improves Rank-1 and mAP accuracy by 0.7% and 0.4%. And we can achieve 89.3% and 80.4% on MARS dataset by combining A^p and A^f. With the consistency loss in the training stage, we can find the Rank-1 and mAP accuracy are improved by 0.5% and 0.7% respectively. With all these proposed components, we improve the Rank-1 and mAP accuracies from 87.8% and 78.0% to 89.8% and 81.1% respectively.

Qualitative results. We visualize the qualitative results in Figure 6 with Grad-CAM [61], which is popularly used in computer vision problem for a visual explanation. The class activation maps (CAMs) generated by the baseline model (second row) and our proposed method (third row) are provided. For the video sequence fragment, in the third and fourth image, the baseline model is incapable of capturing the informative person cues, and our proposed method is more robust to the clutter background.

Retrieval Results As illustrated in Figure 7, we provide the retrieval results on MARS dataset for the baseline model and our proposed method. The illustration presents the improvement of our proposed method.

V. CONCLUSION

This paper proposes an innovative graph representation learning approach for video person Re-ID. The proposed method can learn an adaptive structure-aware adjacency graph over the spatial person regions. By aggregating the contextual messages from neighbors for each node, the intrinsic affinity structure information among person feature nodes is captured adaptively, and the complementary contextual information is further propagated to enrich the person feature representations. Furthermore, we propose a novel regularization to enforce

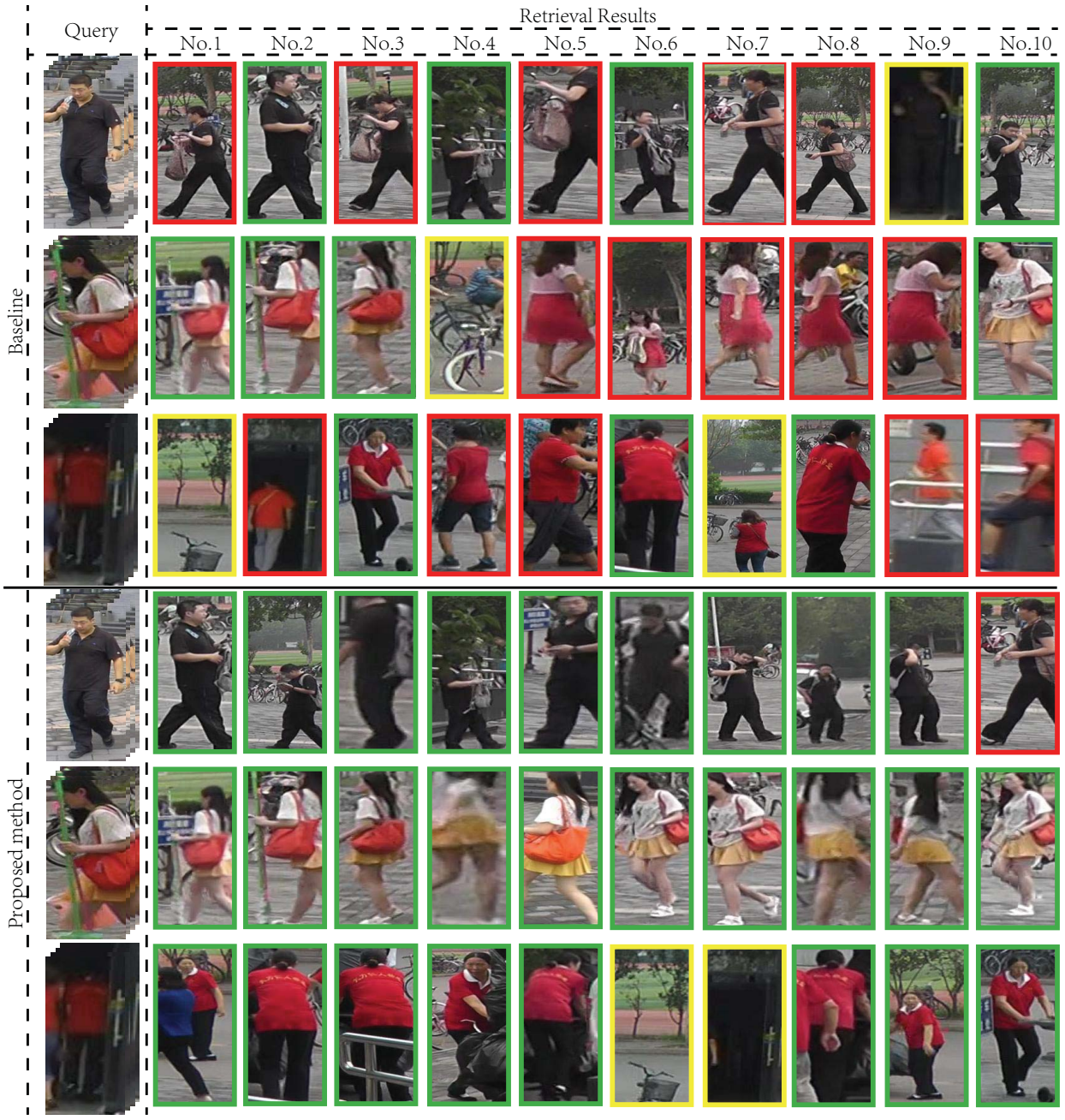

Fig. 7. Comparison of Rank-10 of our proposed method and baseline model. In each row, the images present the videos in the gallery. The images with the green box match the query, and the red box is the wrong matched result. Besides, the images with the yellow box are distracters, which is neglected while calculating the accuracy. **Best viewed in color.**

the consistency among different temporal resolutions, and it is beneficial for learning the compact and discriminative representations. The experimental results on four standard benchmarks demonstrate the effectiveness of the proposed scheme, and extensive ablation studies validate the feasibility of components in the network.

REFERENCES

[1] W. Zajdel, Z. Zivkovic, and B. Krose, "Keeping track of humans: Have i seen this person before?" in *Proc. IEEE Conf. Robot. and Auto.* IEEE, 2005, pp. 2081–2086.

[2] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2016.

[3] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, July 2017.

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

[5] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, July 2017.

[6] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE*

Comput. Soc. Conf. Comput. Vis. Pattern Recognit., June 2018.

[7] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2018.

[8] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2018.

[9] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1366–1377, 2018.

[10] L. Wu, Y. Wang, L. Shao, and M. Wang, "3-d personvlad: Learning deep global representations for video-based person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, 2019.

[11] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, and L. Lin, "Scan: Self-and-collaborative attention network for video person re-identification," *IEEE Trans. Image Process.*, 2019.

[12] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Trans. Image Process.*, 2019.

[13] Y. Liu, Z. Yuan, W. Zhou, and H. Li, "Spatial and temporal mutual promotion for video-based person re-identification," in *Proc. AAAI*, 2019.

[14] J. Li, S. Zhang, and T. Huang, "Multi-scale 3d convolution network for video based person re-identification," in *Proc. AAAI*, 2019.

[15] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *Proc. AAAI*, 2019.

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

[17] A. Borgia, Y. Hua, E. Kodirov, and N. Robertson, "Gan-based pose-aware regulation for video-based person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput.* IEEE, 2019, pp. 1175–1184.

[18] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9317–9326.

[19] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proc. AAAI*, 2018, pp. 7347–7354.

[20] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia*, 2018.

[21] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2018.

[22] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3219–3228.

[23] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conference Comput. Vis.*, September 2018.

[24] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI*, 2019.

[25] F. Zheng, X. Sun, X. Jiang, X. Guo, Z. Yu, and F. Huang, "Pyramidal person re-identification via multi-loss dynamic training," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019.

[26] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

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

[28] J. Zhou, P. Yu, W. Tang, and Y. Wu, "Efficient online local metric adaptation via negative samples for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2420–2428.

[29] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, July 2017.

[30] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 420–429.

[31] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conference Comput. Vis.*, 2018, pp. 650–667.

[32] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3960–3969.

[33] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *Proc. Eur. Conference Comput. Vis.* Springer, 2016, pp. 701–716.

[34] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[35] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, 2009.

[36] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1024–1034.

[37] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[38] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, accepted as poster. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[39] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4800–4810.

[40] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proc. AAAI*, 2018.

[41] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI*, 2018.

[42] D. Cheng, Y. Gong, X. Chang, W. Shi, A. Hauptmann, and N. Zheng, "Deep feature learning via structured graph laplacian embedding for person re-identification," *Pattern Recognit.*, vol. 82, pp. 94–104, 2018.

[43] A. Barman and S. K. Shah, "Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1115–1124.

[44] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conference Comput. Vis.*, 2018, pp. 486–504.

[45] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8649–8658.

[46] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2019.

[47] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5142–5150.

[48] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proc. Eur. Conference Comput. Vis.*, September 2018.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2016.

[50] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3693–3702.

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

[52] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2019, pp. 0–0.

[53] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conference Comput. Vis.* Springer, 2014, pp. 688–703.

[54] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandinavian Conf. Image Anal.* Springer, 2011, pp. 91–102.

[55] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conference Comput. Vis.* Springer, 2016, pp. 868–884.

[56] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV Workshop*, 2016.
[57] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, June 2018.
[58] J. B. Diederik P. Kingma, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[59] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.
[60] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng, "Video-based person re-identification with accumulative motion context," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2788–2802, 2017.
[61] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.