Temporal Extension Topology Learning for Video-based Person Re-Identification

Jiaqi Ning[1], Fei Li[1], Rujie Liu[1], Shun Takeuchi[2], and Genta Suzuki[2]

[1] Fujitsu Research & Development Center, Beijing, China
ningjiaqi@fujitsu.com
[2] Fujitsu Research, Kawasaki, Japan

Abstract. Video-based person re-identification aims to match the same identification from video clips captured by multiple non-overlapping cameras. By effectively exploiting both temporal and spatial clues of a video clip, a more comprehensive representation of the identity in the video clip can be obtained. In this manuscript, we propose a novel graph-based framework, referred as Temporal Extension Adaptive Graph Convolution (TE-AGC) which could effectively mine features in spatial and temporal dimensions in one graph convolution operation. Specifically, TE-AGC adopts a CNN backbone and a key-point detector to extract global and local features as graph nodes. Moreover, a delicate adaptive graph convolution module is designed, which encourages meaningful information transfer by dynamically learning the reliability of local features from multiple frames. Comprehensive experiments on two video person re-identification benchmark datasets have demonstrated the effectiveness and state-of-the-art performance of the proposed method.

Keywords: person ReID · graph convolution network.

1 Introduction

Person re-identification (ReID) [1, 2] is an efficient computer vision technique to retrieve a specific person from multiple non-overlapping cameras. Person ReID has a wide range of applications such as security, video surveillance, etc., and has received extensive attention from researchers. Although many research results have been achieved, this task is still challenging due to background disturbances, occlusions, perspective changes, pose changes and other problems.

There are generally two kinds of person ReID processing methods [1]. One is image-based methods [3–9], which exploit temporally incoherent static images to retrieve pedestrians. The other is video-based methods [10–14], where both training and test data consist of temporally continuous image sequences. In recent years, impressive progress has been made in image-based person ReID. Some practical solutions are proposed especially for complex problems, such as occlusion [9, 15, 16]. However, the information contained in a single image is limited. If the information contained in a short video clip of a pedestrian could be effectively mined, it would significantly benefit the robustness of retrieval

results. Therefore, video-based person ReID methods concurrently utilize spatial and temporal information of the video clip and have the potential to better solve the difficult problems in person ReID.

There have been several typical video-based person ReID methods that aggregate temporal and spatial cues of video clips to obtain discriminative representations. Several elementary methods extract global features from each frame independently. Then the features of each frame are aggregated into the representation of a video clip by a temporal pooling layer or recurrent neural network (RNN) [17–19]. Due to the problems such as occlusion and background noise, these methods usually do not achieve excellent results. Recent works began to focus on the role of local features. Some works divide video frames into rigid horizontal stripes or utilize an attention mechanism to extract local appearance features [20–24]. However, it is hard to align local features learned from videos precisely. Some methods adopt pose estimation model to detect key points of identity in order to obtain well aligned local features [15, 25, 26]. Nevertheless, the noise will be introduced into the extracted local features due to occlusion and inaccurate key point detection. Some works use the Graph Convolution Network (GCN) technique to enhance the description of local features by setting local features as the nodes of GCN [27–29]. In the GCN, information could be transferred between nodes through edges, and the information of nodes can be enhanced or supplemented. However, in the occluded regions, the features are often unintelligible [9]. If all the local parts are considered to have the same reliability for information transmission, it brings in more noise and is terrible for extracting discriminative representations.

In most of graph-based video-based person ReID methods, the transfer metric among nodes is determined by the affinity between feature pairs. This may result in ignoring global contextual information from all other nodes and only considering undirected dependency [14, 27, 28]. Some of them utilize more than one graph to realize temporal and spatial dimension information extraction, which increases the complexity of the method [28, 29, 14].

A novel Temporal Extension Adaptive Graph Convolution (TE-AGC) framework is proposed for video-based person ReID in this manuscript. TE-AGC extracts global and local semantic features from multiple images as graph nodes. Then, the TE-AGC learns the reliability of each local feature extracted from multiple images, encouraging high-reliability nodes to transfer more information to low-reliability nodes, and inhibiting information passing of low-reliability nodes. Further, TE-AGC considers the dependencies of body parts within a frame or across different frames in both temporal dimension and spatial dimension using only one single graph. This way, it could mine comprehensive and discriminative features from the video clip by performing the designed graph convolution. The validity of the TE-AGC is revealed by the experiments on two benchmark datasets, MARS and DukeMTMC-VideoReID.

The main contributions of this paper are as follows: (1) A novel Temporal Extension Adaptive Graph Convolution (TE-AGC) framework for video-based person ReID is proposed. (2) We learned the reliability of local features and adap-

tively pass information from more meaningful nodes to less meaningful nodes. (3) We mine the temporal and spatial dimension information of video clips with one convolution graph.

2 Related Work

2.1 Image-based person ReID

There have been many works for image-based person ReID [3–9]. Benefit from the continuous advance of deep learning technology, the rank-1 accuracy of most image-based person ReID methods on the benchmark dataset is higher than that of human beings [1]. With utilizing the local semantic features and attention mechanisms [16, 30], the performance of person ReID is further improved. In recent years, more researchers have paid attention to the occlusion problem of ReID and achieved fruitful results [9, 15, 16].

2.2 Video-based person ReID

Video-based person ReID can extract richer spatial-temporal clues than image-based person ReID and is expected for more accurate retrieval [10]. Some works extract features for each image of the video clip then aggregate them using temporal pooling or RNN [17–19]. To learn robust representation against pose changes and occlusions, the local semantic features and attention mechanisms are also be used in the video-based person ReID to improve the performance [20–26, 15]. Different from the image-based person ReID, the time dimension is added, and both spatial attention and temporal attention are used to mine the information of a video clip.

2.3 Graph Convolution

Graphs are often used to model the relationship between different nodes. Graph Convolution Network (GCN) simply utilizes the convolution operation of image processing in graph structure data processing for the first time [31]. Great success has been achieved in many computer vision tasks, like skeleton-based action recognition [32], object detection [33] and person ReID [9, 27–29]. Some methods have been proposed to use the GCN in person ReID. Some treat the image as nodes of graph, ignoring the relationship between different body parts within or across frames. In addition, some recent works model the temporal and spatial relationships of nodes in two or more graphs. For example, the Spatial-Temporal Graph Convolutional Network (STGCN) [28] constructs two graph convolution branches. The spatial relation of human body and the temporal relation from the adjacent frames are learned in two different graphs.

3 Method

We aim to develop an efficient spatial-temporal representation for video-based person ReID. To this end, the Temporal Extension Adaptive Graph Convolution (TE-AGC) framework is proposed in this manuscript, as shown in Fig. 1. In general, the whole frame contains two parts. One is to extract preliminary semantic features, and the other is to obtain an improved discriminative representation of a video clip.

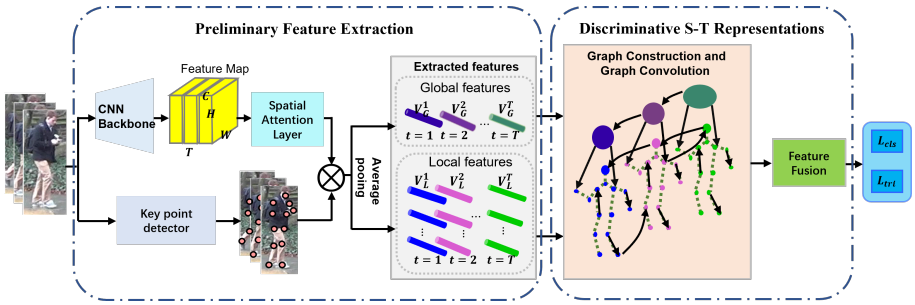

Fig. 1. The overall architecture of the proposed TE-AGC. It includes a backbone network, a spatial attention layer and a key-point detector to extract the global and local semantic features. Moreover, it also includes a graph construction and graph convolution layer and a feature fusion block to obtain discriminative spatial-temporal representation for each video clip.

3.1 Semantic Feature Extraction

First, we perform preliminary semantic feature extraction. It has been demonstrated that part features are effective for person ReID [1, 2]. Inspired by this idea, we aim to extract both global and local semantic features in this module. To better resist the viewpoint variation and misalignment, a key-point detector is utilized to locate key points. Then we extract local features from different key points. It should be noted that, although human key-point detection is a relative mature technique, there still exist key point position errors and key point confidence errors in some cases [9]. Thus, the module introduced in **3.2** is necessary and will enhance the features.

Given a video clip, we randomly sample T frames. These randomly sampled frames are denoted as $\{I^t\}_{t=1}^{T}$ and t is the index of the frame. The backbone network is used to generate the initial feature maps for each frame. Then we use a spatial attention layer to enhance the spatial feature and suppress the interference information. The spatial attention layer is implemented by a convolutional layer followed by a sigmoid activation function. Then the initial feature maps are weighted by the attention layer. The set of feature maps after spatial attention

operation are denoted as $\mathcal{F} = \{F^t\}_{t=1}^T$, where $F^t \in \mathbb{R}^{C\times H\times W}$ and H, W, C denote the height, width and channel number respectively.

We use a key point detector to help extracting aligned local features. For each frame, the number of extracted key point is K. The local and global semantic features of each frame are computed as follows. For distinction, we denote the features at this stage as V and denote the features output by the module introduced in 3.2 as $V^{'}$.

$$V_L{}^t = \{v_k^t\}_{k=1}^K = \{g_{GAP}\left(F^t \bigotimes m_k^t\right)\}_{k=1}^K \tag{1}$$

$$V_G{}^t = g_{GAP}\left(F^t\right) \tag{2}$$

where $V_L{}^t \in \mathbb{R}^{K\times C}$denotes the local features of the frame t, which include semantic local features of K key points. $v_k^t \in \mathbb{R}^{1\times C}$is the local feature around k^{th} key point of frame t. m_k^t is derived from the heatmap of the k^{th} key point of frame t by normalizing original heatmap with a SoftMax function. $g_{GAP}(\cdot)$ refers to global average pooling operation. $\bigotimes$ is element by element multiplication operation. $V_G{}^t \in \mathbb{R}^{1\times C}$ denotes the global feature of the frame t. Therefore, the preliminary semantic of this video clip contains local feature $\{V_L{}^t\}_{t=1}^T$ and global feature $\{V_G{}^t\}_{t=1}^T$.

3.2 Temporal extension adaptive graph convolution layer

After extracting each frame's preliminary global and local features, we employ advanced GCN to mine spatial-temporal representation from video frames.

A graph convolution can be operated as [31]:

$$O = \hat{A}XW \tag{3}$$

where $\hat{A}$ is normalized version of the adjacent matrix A, and X is the feature matrix which contains features of all nodes. W refers to parameters to be learned. O is the output after graph convolution operation.

In our method, the preliminary global feature and local feature of each frame within one video clip are treated as the graph nodes. For a video clip, the number of nodes is $N = T \times (K + 1)$ including features of T frames, K local features and 1 global feature per frame. The feature matrix X with the size of $N \times C$ is constructed by the concatenation of $\{V_L{}^t\}_{t=1}^T$ and $\{V_G{}^t\}_{t=1}^T$ in vertical direction. The adjacent matrix $A \in \mathbb{R}^{N\times N}$ illustrates the topology of the graph. $A(i, j)$ represents the information propagation metric from node j to node i. When $A(i, j)$ equals to zero, no information transfers from node j to node i.

Most of the GCN-based person ReID methods obtain the $A(i, j)$ by calculating the feature affinity between node j and node i. However, this calculation method might introduce noise when some nodes are unreliable. Considering that, we propose a Temporal Extension Adaptive Graph Convolution (TE-AGC) layer to calculate the adjacent matrix A.

Our method efficiently utilizes one single graph to transmit information in both temporal and spatial dimensions. In addition to local features, we introduce global features of each frame as graph nodes, considering complex spatial-temporal dependencies of different body parts and the whole body within a frame or across frames. Linkage modes between nodes are shown in fig. 2. They include local features within single frame (① in fig. 2), local features and global features within single frame (④ in fig. 2), corresponding local features across frames (② in fig. 2), non-corresponding local features across frames (③ in fig. 2) and global features across frames (⑤ in fig. 2). It should be noted that we just illustrate the modes of connections and do not draw all the connections in fig.2. The connections between different local features including both within and across frames are defined by human skeleton. Information transfer will be performed among key points in adjacent positions.

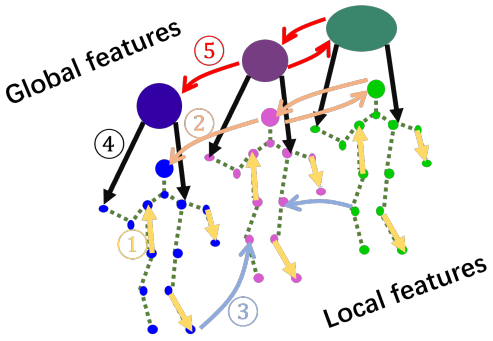

Fig. 2. Linkage modes among nodes.

After determining whether there is a connection relationship between each node, our method delicately designs the information propagation metric. Inspired by the assumption proposed in [9] that the meaningful local feature is more similar to the global feature than the meaningless local feature, a method to learn the reliability of local features and compute the value of $A(i,j)$ is proposed. Suppose node i is local feature of k^{th} key point of frame t, the reliability of node i, referred as D_i is learned as follows:

$$D_i = FC\left(BN\left(abs\left(v_k^t - {V_G}^t\right)\right)\right) \tag{4}$$

where $abs(\cdot)$ and $BN(\cdot)$ are absolution and batch normalization. $FC(\cdot)$ is fully connected layer mapping a vector with size $1 \times C$ to a real number. The reliabilities of local features are normalized by a SoftMax operation.

If both node j and node i are local features and exist information transfer (types ① ② ③ in Fig. 2), $A(i,j)$ is calculated as follow:

$$A(i,j) = ReLU(1 + \alpha(D_j - D_i)) \times \alpha(D_i + D_j) \tag{5}$$

where α is a hyperparameter larger than zero to balance the transfer metric between global features and local features. D_i and D_j are the reliability of node i and node j calculated by (4). Obviously, if node j is more reliable than node i, node j will transmit more information to node i. If node j is global feature and node i is local feature and exist information transfer (type ④ in Fig. 2), the transfer metric is set as $ReLU(1-\beta \times D_i)$, where β is another hyperparameter. And the pass-through metrics between global feature and global feature (type ⑤ in Fig. 2) is set as $1/(T-1)$. The matrix A is normalized as $\hat{A}$ by applying $L1$ normalization operation to each row of A.

By (3), output feature O after the graph convolution can be obtained. For stabilize training, we fuse O and the input features X as in the ResNet [34]. The output of the TE-AGC layer is the improved feature $V^{'}$.

$$V^{'} = FC(X) + ReLU(O) \tag{6}$$

where $ReLU(\cdot)$ is the activation function that if the input is greater than zero, it remains unchanged; otherwise the output is zero. The $V^{'}$ contains improved local features $\{V_L^{'t}\}_{t=1}^{T}$ and improved global features $\{V_G^{'t}\}_{t=1}^{T}$. The $V_L^{'t} = \{v^{'t}_k\}_{k=1}^{K}$, $v^{'t}_k$ is the improved local feature around k^{th} key point of frame t.

With this adaptive method we proposed, the linkage among nodes is decided by the input features. Nodes with high reliability will transfer more information to nodes with low reliability. Therefore, the information can be transmitted more effectively through graph convolution.

3.3 Model Optimizing

After obtaining the improved features and preliminary semantic features, we will further incorporate the global and local representations. We employ a temporal average pooling layer $g_{TAP}(\{\cdot\}_{t=1}^{T})$ to generate the time average (TA) feature vector

$$V_G{}^{TA} = g_{TAP}(\{V_G{}^{t}\}_{t=1}^{T}) \tag{7}$$

$$V_G^{'TA} = g_{TAP}(\{V_G^{'t}\}_{t=1}^{T}) \tag{8}$$

And we obtain the local and global combined TA feature vector by

$$V_C{}^{TA} = g_{TAP}(\{\sum_{k=1}^{K} v_k^t\}_{t=1}^{T}) + V_G{}^{TA} \tag{9}$$

$$V_C^{'TA} = g_{TAP}(\{\sum_{k=1}^{K} v^{'t}_k\}_{t=1}^{T}) + V_G^{'TA} \tag{10}$$

The model is optimized by lose function. We utilized identification loss and triplet loss for $V_G^{'TA}$, $V_C^{'TA}$, $V_G{}^{TA}$ and $V_C{}^{TA}$. We combine the identification loss

and triplet loss as the total loss with a weighted parameter λ to balance weights of different kind of loss. We adopt triplet loss with hard mining strategy [35] and identification loss with label smoothing regularization [36] to optimize the loss function.

4 Experiments

4.1 Dataset and Implementation

Datasets. Two benchmarks of video-based person ReID datasets, MARS and DukeMTMC-VideoReID, are utilized to evaluate the TE-AGC. MARS, the largest video-based person ReID dataset, contains 17503 tracklets from 1261 identities and 3248 distractor sequences. 625 identities are contained in training set and 636 identities are contained in test set. DukeMTMC-Video is derived from the DukeMTMC dataset, with 4832 tracklets from 1812 identities. There are 408, 702, 702 identities for distraction, training and testing respectively.

Evaluation protocols. We adopt the mean average precision (mAP) and the Cumulative Matching Characteristic (CMC) to evaluate the performance of our method.

Implementation Details. We set $T = 3$, which means we randomly select three frames as an input sample from a variable-length video clip. Each image is resized to 256×128 pixels. Random horizontal flips and random erasing are performed in the image augmentation process. We employ the ResNet-50 [34] pre-trained on ImageNet [37] as the backbone network after removing the global average pooling and full connected layers. We use HR-Net [38] pretrained on the COCO dataset [39] as the human key points detector.In our method, 13 body key-points are used. During the training period, the learning rate is initialized as 3.5×10^{-4} and decayed by 5 after every 70 epochs. The optimizer is Adam with weight decay 5×10^{-4}. The model is totally trained for 500 epochs. During inference, the representation of a video clip to calculate the similar scores is the $V_C^{'TA}$. It should be noted that, when we set $T = 4$ or more, the ReID performance is very similar to $T = 3$. Considering the calculation cost, $T = 3$ is suitable for our method. Therefore, the following experimental analysis is completed under the setting of $T = 3$.

4.2 Comparison with state-of-the-arts

Table 1 makes a comparison between our method and state-of-the-art algorithms on MARS and DukeMTMC-VideoReID datasets. Our method has achieved state-of-the-art performance.

Results on MARS. Our method is compared with 12 state-of-the-art methods on MARS dataset. Among these methods, AGRL, STGCN and MGH are three other graph-based methods. Compared with these graph-based methods, our method achieves higher Rank-1, Rank-5 accuracy and mAP. There are two main reasons for this improvement. One is our method considers the complex

spatial-temporal relation among different body parts and whole body within a frame or across frames. On the other hand, instead of using pair-wise feature affinity, the information pass metrics we designed encourage reliable nodes to pass more information to other modes.

Results on DukeMTMC-Video. Our method is compared with 11 state-of-the-art methods on DukeMTMC-Video dataset. Our method has gotten 97.2% rank-1 results and 96.3% mAP, which exceeds the vast majority of state-of-the-art methods. The comparison verifies the effectiveness of our method.

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

Methods	MARS			DukeMTMC-VideoReID		
	Rank-1	Rank-5	mAp	Rank-1	Rank-5	mAP
STA[20]	86.3	95.7	80.8	96.2	99.3	94.9
GLTR [40]	87.0	95.8	78.5	96.3	99.3	93.7
COSAM [23]	84.9	95.5	79.9	95.4	99.3	94.1
VRSTC [41]	88.5	96.5	82.3	95.0	99.1	93.5
RGSAT [42]	89.4	96.9	84.0	97.2	**99.4**	95.8
AGRL [27]	89.8	96.1	81.1	96.7	99.2	94.2
TCLNet [24]	89.8	-	85.1	96.9	-	96.2
STGCN [28]	90.0	96.4	83.7	**97.3**	99.3	95.7
MGH [29]	90.0	96.7	85.8	-	-	-
AP3D [43]	90.1	-	85.1	96.3	-	95.6
AFA [44]	90.2	96.6	82.9	97.2	**99.4**	95.4
BiCnet-TKS [13]	90.2	-	**86.0**	96.3	-	96.1
TE-AGC (ours)	**90.7**	**97.5**	85.8	97.2	**99.4**	**96.3**

4.3 Model Component Analysis

The architecture of our method has only one branch, which is both concise and effective. Here the contribution of each part of TE-AGC is evaluated and results on MARS dataset are reported.

Table 2 reports the experimental results of the ablation studies for TE-AGC. The 1^{st} line can be regarded as baseline result of our method. In the 1^{st} line, for each frame, we use the backbone network and key point detector to get global and local features. Then we combine all features like (7) and (9). Compared with 1^{st} line, the 2^{nd} line shows the result of adding the Spatial Attention (SA) layer. In addition, the 3^{rd} line adds GCN layer we designed but does not use the spatial attention layer, which means only removing the spatial attention layer compared with the entire architecture (the 4^{th} line).

Comparing the 1^{st} line with the 2^{nd} line or comparing the 3^{rd} line with the 4^{th} line, the effectiveness of the spatial attention can be directly proved. Though it is

Table 2. Component analysis of the effectiveness of each component of TE-AGC on MARS dataset.

Number	Methods	MARS		
		Rank-1	Rank-5	mAP
1	TE-AGC –GCN –SA	88.7	96.1	83.8
2	TE-AGC –GCN	89.0	96.5	84.1
3	TE-AGC –SA	90.3	96.9	85.3
4	TE-AGC	**90.7**	**97.5**	**85.8**

relatively simple, it still has a certain inhibitory effect on background noise. The comparison of the 2^{nd} & 4^{th} lines and the comparison of the 1^{st} & 3^{rd} lines show the effect of the graph convolution layer we designed. The information of each node is effectively transferred and enhanced during the graph convolution. Finally, the powerful spatio-temporal representation for the video is obtained.

5 Conclusion

In this paper, a novel Temporal Extension Adaptive Graph Convolution (TE-AGC) framework is proposed for video-based person ReID. The TE-AGC could mine features in spatial and temporal dimensions in one graph convolution operation effectively. The TE-AGC utilizes a CNN backbone, a simple spatial attention layer and a key-point detector to extract global and local features. A delicate adaptive graph convolution module is designed to encourage meaningful information transfer by dynamically learning the reliability of local features from multiple frames. We combine the global feature and local features of each frame with a future fusion module to obtain discriminative representations of each video clip. The effectiveness of the TE-AGC method is verified by a large number of experiments on two video datasets.

References

1. Ye, M., Shen, J., Lin, G., Xiang, T., Hoi, S.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP** (2021) 1–1
2. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. (2016)
3. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proc IEEECONFERENCE on Computer Vision & Patternrecognition. (2010) 2360–2367
4. Liu, C., Gong, S., Chen, C.L., Lin, X.: Person re-identification: What features are important? Springer, Berlin, Heidelberg (2012)

5. Liao, S., Yang, H., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
6. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
7. Gou, M., Fei, X., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Computer Vision–ECCV 2014. (2014)
8. Zheng, W.S., Xiang, L., Tao, X., Liao, S., Lai, J., Gong, S.: Partial person re-identification. In: IEEE International Conference on Computer Vision. (2016)
9. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J.: High-order information matters: Learning relation and topology for occluded person re-identification. (2020)
10. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV). (2017)
11. Chen, D., Li, H., Tong, X., Shuai, Y., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
12. Xu, S., Yu, C., Kang, G., Yang, Y., Pan, Z.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV). (2017)
13. Hou, R., Chang, H., Ma, B., Huang, R., Shan, S.: Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. (2021)
14. Liu, J., Zha, Z.J., Wu, W., Zheng, K., Sun, Q.: Spatial-temporal correlation and topology learning for person re-identification in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2021) 4370–4379
15. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). (2019)
16. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
17. Mclaughlin, N., Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Computer Vision & Pattern Recognition. (2016)
18. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person reid. (2018)
19. Liang, Z., Zhi, B., Sun, Y., Wang, J., Qi, T.: Mars: A video benchmark for large-scale person re-identification. In: European Conference on Computer Vision. (2016)
20. Fu, Y., Wang, X., Wei, Y., Huang, T.S.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: National Conference on Artificial Intelligence. (2019)
21. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 369–378
22. Ouyang, D., Zhang, Y., Shao, J.: Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks. Pattern Recognition Letters (2018) S0167865518301752

23. Subramaniam, A., Nambiar, A., Mittal, A.: Co-segmentation inspired attention networks for video-based person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 562–572
24. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Temporal complementary learning for video person re-identification. In: European conference on computer vision, Springer (2020) 388–405
25. Jones, M.J., Rambhatla, S.: Body part alignment and temporal attention for video-based person re-identification. In: BMVC. (2019)
26. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1077–1085
27. Wu, Y., Bourahla, O.E.F., Li, X., Wu, F., Tian, Q., Zhou, X.: Adaptive graph representation learning for video person re-identification. IEEE Transactions on Image Processing **29** (2020) 8821–8830
28. Yang, J., Zheng, W.S., Yang, Q., Chen, Y.C., Tian, Q.: Spatial-temporal graph convolutional network for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 3289–3299
29. Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y., Shao, L.: Learning multi-granular hypergraphs for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 2899–2908
30. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. (2020) 3186–3195
31. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
32. Obinata, Y., Yamamoto, T.: Temporal extension module for skeleton-based action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE (2021) 534–540
33. Shi, W., Rajkumar, R.: Point-gnn: Graph neural network for 3d object detection in a point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 1711–1719
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
35. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2818–2826
37. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee (2009) 248–255
38. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 5693–5703
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755

40. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 3958–3967
41. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Vrstc: Occlusion-free video person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 7183–7192
42. Li, X., Zhou, W., Zhou, Y., Li, H.: Relation-guided spatial attention and temporal refinement for video-based person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 11434–11441
43. Gu, X., Chang, H., Ma, B., Zhang, H., Chen, X.: Appearance-preserving 3d convolution for video-based person re-identification. In: European Conference on Computer Vision, Springer (2020) 228–243
44. Chen, G., Rao, Y., Lu, J., Zhou, J.: Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In: European Conference on Computer Vision, Springer (2020) 660–676