Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification

Yingquan Wang[1], Pingping Zhang[2], Shang Gao[3], Xia Geng[1], Hu Lu[1]*, Dong Wang[3]
[1] School of Computer and Communication Engineering, Jiangsu University
[2] School of Artificial Intelligence, Dalian University of Technology
[3] School of Information and Communication Engineering, Dalian University of Technology
{yingquan1995, gs940601k}@gmail.com; {zhpp, wdice}@dlut.edu.cn; {Luhu, Gengxia}@ujs.edu.cn

Abstract

Video-based person re-identification aims to associate the video clips of the same person across multiple non-overlapping cameras. Spatial-temporal representations can provide richer and complementary information between frames, which are crucial to distinguish the target person when occlusion occurs. This paper proposes a novel Pyramid Spatial-Temporal Aggregation (PSTA) framework to aggregate the frame-level features progressively and fuse the hierarchical temporal features into a final video-level representation. Thus, short-term and long-term temporal information could be well exploited by different hierarchies. Furthermore, a Spatial-Temporal Aggregation Module (STAM) is proposed to enhance the aggregation capability of PSTA. It mainly consists of two novel attention blocks: Spatial Reference Attention (SRA) and Temporal Reference Attention (TRA). SRA explores the spatial correlations within a frame to determine the attention weight of each location. While TRA extends SRA with the correlations between adjacent frames, temporal consistency information can be fully explored to suppress the interference features and strengthen the discriminative ones. Extensive experiments on several challenging benchmarks demonstrate the effectiveness of the proposed PSTA, and our full model reaches 91.5% *and* 98.3% *Rank-1 accuracy on MARS and DukeMTMC-VID benchmarks. The source code is available at* `https://github.com/WangYQ9/VideoReID-PSTA`.

1. Introduction

Person re-identification (ReID) aims to match a particular person from non-overlapping camera views, which is an important technology in many applications, such as video surveillance, tracking, and smart city. However, it is challenging due to many practical obstacles, such as background clutter, blur, occlusion and viewpoint variations.

Recently, image-based person ReID has achieved impressive progress [2, 39, 54, 26, 43]. Most of these works focus on extracting more discriminative features within a single image. Therefore, it is difficult to retrieve the target person when occlusion occurs, or missing crucial parts. In contrast, the richer spatial-temporal information can alleviate the limitation of image-based ReID and is more powerful to obtain discriminative features and robust results.

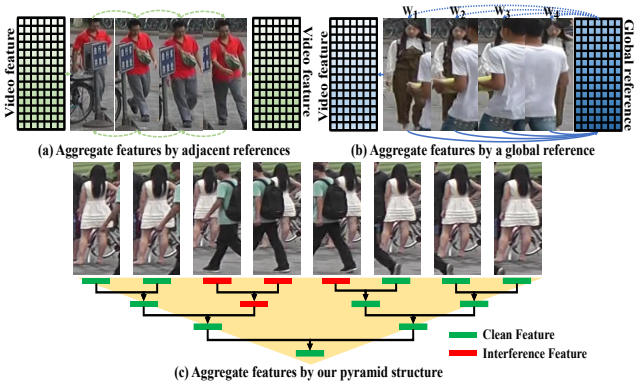


Figure 1. Illustration of various solutions that employ temporal context information to aggregate frame-level features. (a) Transferring information from adjacent frames. (b) Using a global reference to guide the attention of each frame. (c) Fusing frame-level features with a pyramid structure (ours). The green lines indicate the clean features of the target person while the red ones indicate the features interfered by occlusions. The pyramid structure can alleviate the irrelevant feature by aggregating progressively.

Several works [17, 38, 49, 32] have been proposed to enhance the discriminative features of the target person and suppress the irrelevant features with the help of spatial-temporal context information. Subramaniam *et al.* [38] propose a co-segmentation module to activate a common set of features across multiple frames. Hou *et al.* [17] aim to solve the problem of partial occlusion. They use the information from adjacent frames to reconstruct the occlusion part, thus could alleviate the interference of irrelevant features. The message passing flow of these methods is shown in Fig. 1 (a). Although the information from the temporal adjacent frames can somewhat suppress the occluded features, it would lose efficacy when long-range occlusion occurs because it lacks the long-term dependence. Yan *et*

*Corresponding Author

al. [53] propose to learn the attention from a global view by constructing a global reference. The message passing flow of these methods is shown in Fig. 1 (b). Although the reference can capture global information, there is no guarantee that it can represent the target person well. For example, this method may focus on the occlusion parts when most of the frames are occluded. Jiang *et al.* [21] propose to infer attentions of frame-level features by constructing a relation embedding for each pair of frames in the tracklet. This kind of method fully exchanges the message among the whole tracklet. However, it would suffer from the high redundancy among frames and the large distribution space for relation embedding. On the other hand, it is not computationally efficient to construct all relations of frames.

To alleviate the problems mentioned above, we propose a novel Pyramid Spatial-Temporal Aggregation (PSTA) framework for high-performance video-based person ReID. Fig. 1 (c) illustrates our basic idea. The adjacent frame-level feature maps are grouped into pairs and then sent into a hierarchical aggregation module. As the process goes on, long-term dependence could be constructed in the later stage without losing the short-term information from the previous stage. It is obvious that after aggregating by our PSTA, the proportions of the clean features are increased. Therefore, the key to improving performance is how to aggregate the adjacent features, such that the fused features can be more discriminative and with less interference from occlusions. We argue that a well-designed aggregation module should satisfy two requirements: 1) foreground features can be strengthened with the intra-frame information. 2) features of the target person can be enhanced, and the non-target information can be suppressed with the inter-frame correlations. We propose a Spatial-Temporal Aggregation Module (STAM) according to the above two requirements. More specifically, it consists of two key components: Spatial Reference Attention (SRA) and Temporal Reference Attention (TRA). SRA explores the spatial correlations within a frame to determine the attention weight of each location. While TRA extends SRA with the correlations between adjacent frames, such that temporal consistency information can be fully explored to suppress the interference features and strengthen the discriminative ones.

In summary, the main contributions are as follow:

- We propose a novel Pyramid Spatial-Temporal Aggregation (PSTA) framework to aggregate the frame-level features step-by-step, establishing the long-term dependence while maintaining the short-term information effectively and efficiently.
- We propose a novel feature aggregation module (STAM) which considers both intra-frame and inter-frame correlations to suppress the interference features and enhance the discriminative ones.
- Extensive experiments demonstrate that our PSTA achieves state-of-the-art performance on several video-based person ReID benchmarks. Our full model reaches 91.5% and 98.3% Rank-1 accuracies on MARS and DukeMTMC-VID benchmarks.

2. Related Work

Comparing with image-based person ReID, video-based person ReID can utilize temporal information to retrieve a person more precisely. To extract video-level features, some works [55, 24, 29] employ temporal pooling across all time stamps. For example, Gao *et al.* [11] apply average pooling to obtain video features. However, simply pooling features may lose much temporal information. Thus, some researchers apply Recurrent Neural Networks (RNNs) for exploring sequence relation [57, 6, 34]. Dai *et al.* [6] first extract image-level features, then utilize two cascade Bi-LSTM networks and temporal pooling to aggregate frame-wise features. However, RNNs may not extract robust temporal information, since Zhang *et al.* [52] prove RNNs can achieve better performance with orderless sampling. To capture spatial-temporal cues directly, 3D Convolution Neural Networks (CNNs) are popular [27, 3]. For example, Carreira *et al.* [1] propose two-stream Inflated 3D CNNs (I3D) for spatial-temporal feature learning in action recognition. Li *et al.* [23] further present a compact Multi-scale 3D CNN (M3D) for video-based ReID. Gu *et al.* [12] build an Appearance-Preserving 3D CNN (AP3D) for handling the appearance destruction problem. Despite their promising performance, these works introduce a mass of parameters and computation. Recently, some works use Graph Convolution Networks (GCNs) and their variants to extract video representations [49, 46]. For example, Yang *et al.* [50] propose a Spatial-Temporal GCN (STGCN) for mining the spatial and temporal relation. However, most of these methods have a complex structure which may cause the model hard to optimize. To handle these issues, we propose a novel yet simple structure, namely the Pyramid Spatial-Temporal Aggregation (PSTA) framework, to aggregate the frame-level features, which establish long-term dependence without losing the useful local information.

Meanwhile, attention-based methods are widely used in video-based person ReID [16, 48, 37, 38]. For example, Fu *et al.* [10] propose a Spatial-Temporal Attention (STA) approach to emphasize discriminative features. Inspired by the self-attention mechanism [42] for machine translation, Wang *et al.* [45] propose the non-local network to mine the long-range spatial and temporal dependencies. To incorporate video characteristics, Liu *et al.* [28] propose a Non-local Video Attention Network (NVAN) by inserting non-local modules into different stages of ResNet-50 [13]. Li *et al.* [22] further employ the dilated convolution to mine the multi-scale temporal cues. Based on Relation-Aware Global Attention (RAGA) [54], Zhang *et al.* [53] propose MG-RAFA to fuse the image-level fea-

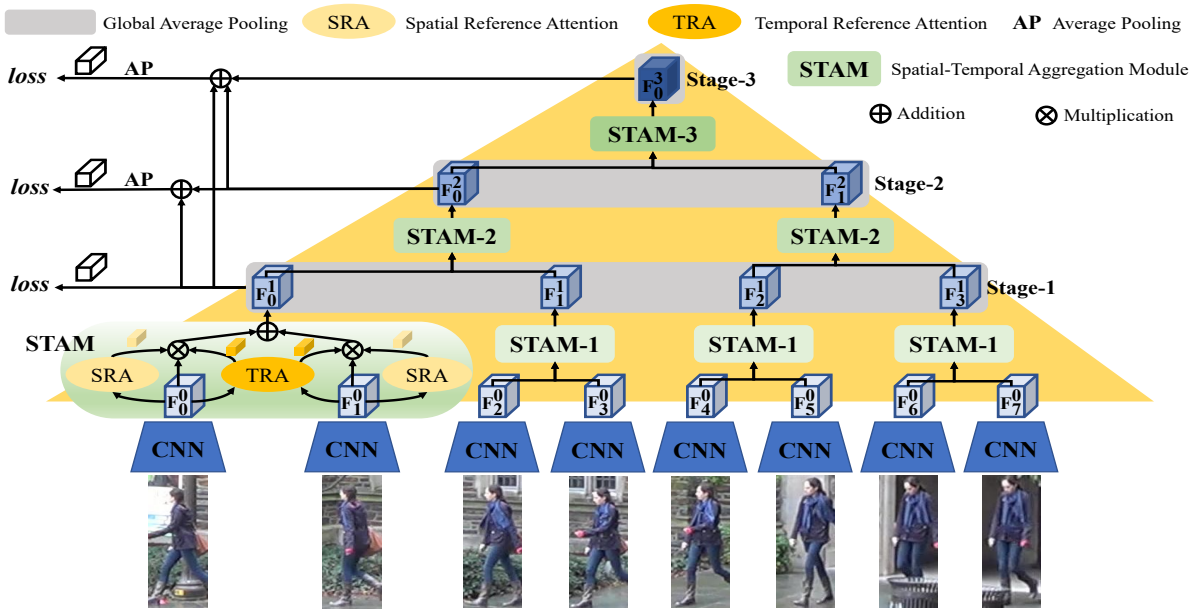


Figure 2. Overall structure of our PSTA framework. Here, we use eight frames ($T = 8$) as an example. Notably, the Spatial-Temporal Aggregation Module (STAM) has the same structure in the pipeline, and the parameters of STAM are **shared** in the same stage.

tures. However, MG-RAFA may lose some local temporal information. Moreover, transformer-based methods become popular in computer vision [8]. Recently, some researchers introduce transformers into the video-based person ReID [51, 30] and achieve promising performance. However, these transformer-based methods may require lots of computation resources, which may increase the difficulty of implementation in the real scenario. In this paper, we propose a Spatial-Temporal Aggregation Module (STAM) to generate the discriminative features from the current and adjacent feature maps with limited computational cost.

3. Proposed Approach

We propose a novel Pyramid Spatial-Temporal Aggregation (PSTA) framework for video-based person ReID. The overall framework is shown in Fig. 2. It aggregates the frame-level features step by step, by which the long-term relation between frames can be established progressively while the short-term dependence can also be fully utilized. Then we propose Spatial-Temporal Aggregation Module (STAM) to aggregate adjacent features further. It enhances the target-relevant features while suppresses the interference feature by the intra-frame attention and the inter-frame attention. In this section, we first introduce the pyramid structure of our PSTA framework. Then the details about each module of STAM are described in Section 3.2. Finally, the loss functions are presented in Section 3.3.

3.1. Pyramid Structure

For a video sequence, we sample a tracklet with T frames denoted as $\mathcal{V} = \{I_1, I_2, \cdots, I_T\}$. As shown in Fig. 2, the tracklet $\mathcal{V}$ is first fed into a feature extractor ($e.g.$, ResNet-50 [13]) to obtain a set of feature maps $\mathcal{F}' = \{F_1^0, \cdots, F_T^0\}$, where $F_t^0 \in \mathbb{R}^{C\times H\times W}$ (C, H and W represent the number of channel, height and width).

Then, a Spatial-Temporal Aggregation Module (STAM) (the details can be found in Section3.2.) takes adjacent feature maps $\{F_{2t-1}^0, F_{2t}^0 | i \in \{1, 2, \cdots, \frac{T}{2}\}\}$ as input and aggregates them into local-temporal features,

$$F_t^n = STAM^n(F_{2t-1}^{n-1}, F_{2t}^{n-1}), t = 1, 2, \cdots, \tfrac{T}{2^n}. \quad (1)$$

Here $n = 1$ means stage of STAM. The progress mentioned above can continue until there is only one output from STAM as shown in Fig. 2. In this way, hierarchical features with different temporal receptive fields can be obtained. For example, an output feature of STAM-2, denoted as F_t^2, contains the information of $2^2 = 4$ frame-level features. Thus long-term dependence can be established as the stage increases without losing the relation information obtained from previous stages.

The output feature set F_t^n is then sent into a Global Average Pooling (GAP) layer to get the stage features F^n

$$F^n = \frac{1}{T'}\frac{1}{W}\frac{1}{H}\sum_{t=1}^{T'}\sum_{w=1}^{W}\sum_{h=1}^{H} f_{t,w,h}^n, \quad (2)$$

where $T' = \frac{T}{2^n}$ is the output number of stage n, $f_{t,w,h}^n \in \mathbb{R}^C$ denotes the vector at the location (w, h) of F_t^n.

In the end, multi-stage features F^n are fused as the final video-level feature

$$X^N = \frac{1}{N}\sum_{n=1}^{N} F^n, \quad (3)$$

where $N = 1, 2, ..., \log_2 T$. To make sure that features at every stage can represent the sequence well, we employ supervision to all stage features. During testing, we use the averaged feature of all stages as the video-level feature.

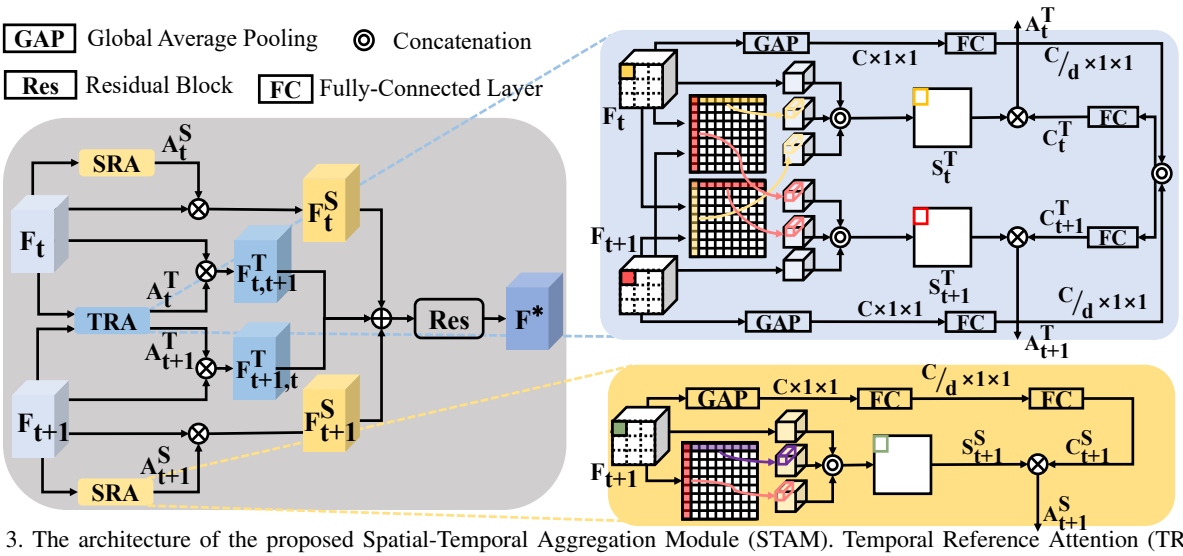


Figure 3. The architecture of the proposed Spatial-Temporal Aggregation Module (STAM). Temporal Reference Attention (TRA) and Spatial Reference Attention (SRA) are shown in the right part.

3.2. Spatial-Temporal Aggregation Module

The attention mechanism has been widely used to obtain the discriminative representation in person ReID. However, most of attention-based modules can not extract inter-frame and intra-frame features simultaneously. To relieve this issue, we design a novel STAM to strengthen the foreground feature by intra-frame information and enhance the target person information by the inter-frame correlations. The architecture of STAM is illustrated in Fig. 3. It consists of two key components: Spatial Reference Attention (SRA) and Temporal Reference Attention (TRA).

Once a pair of temporal adjacent feature maps $\{F_t^n, F_{t+1}^n\}$ is fed into the STAM, it passes through two attention blocks, SRA and TRA, firstly and then the attention maps $A_t^n, A_{t+1}^n, A_{t,t+1}^n, A_{t+1,t}^n$ considering different relations can be obtained. Here A_t^n is the attention map of F_t^n obtained by SRA and $A_{t,t+1}^n$ is the output of TRA with the inputs F_t^n and its reference F_{t+1}^n. Afterward, the refined feature maps can be obtained by

$$F_t^S = A_t \circ F_t, F_{t,t+1}^T = A_{t,t+1} \circ F_t, \tag{4}$$

where $\circ$ is Hadamard product. Then we adopt element-wise addition followed by a residual block [13] to fuse the refined feature which fully explores the discriminative feature of the input clip

$$F^n = Res\Big[(F_{t,t+1}^T + F_{t+1,t}^T) + (F_t^S + F_{t+1}^S)\Big], \tag{5}$$

where the $Res(\cdot)$ is a residual block in [13].

3.2.1 Spatial Reference Attention

The spatial information plays an important role in video-based person ReID. The proposed pyramid aggregation structure can aggregate long-term information appropriately. However, the spatial cues may easily miss out during the feature fusion. To handle this problem, we propose a Spatial Reference Attention (SRA) to enhance the discriminative spatial feature and suppress the interference information. As shown in Eq. 4, directly learning A_i^n can be expensive with a large number of parameters. Inspired by [23], we factorize A_i^n into two low-dimension attention masks as:

$$A_t = S^S \circ C^S, \tag{6}$$

where $S_t^S \in \mathbb{R}^{1\times H\times W}$ and $C_t^S \in \mathbb{R}^{C\times 1\times 1}$ represent the spatial and channel attention masks, respectively. To learn the two attention masks, SRA introduces two branches, shown in the right part of Fig. 3.

Spatial Attention Learning: Inspired by [54], we treat the input feature map F_i as a graph G_s with $N = W \times H$ nodes. Each node corresponds to a $C-$dimensional attribute vector x_i. As shown in Fig. 3, we define the relation from node i to node j as:

$$r_{i,j}^S = \theta^S(x_i)^T \phi^S(x_j), \tag{7}$$

where θ^S and ϕ^S are two embedding functions implemented by a 1×1 convolutional layer followed by a Batch Normalization (BN) [19] layer and ReLU [13] activation function. Then the relation vector can be defined as:

$$r_i^S = \gamma^S\big(r_{i,1}^S \cdots, r_{i,N}^S, r_{1,i}^S \cdots r_{N,i}^S\big), \tag{8}$$

where γ^S is another embedding function which has the same structure with θ^S. Then the relation-value vector v_i^S is constructed by stacking the relation vector r_i^S and its corresponding embedding vector $\beta(x_i)$

$$v_i^S = [r_i^S, \beta(x_i)], \tag{9}$$

where $\beta(x_i)$ is an embedding function. Then, we build the relation-value matrix V^S according to the node index of v_i^S. Finally, the spatial attention map S^S can be obtained by passing the relation-value vector through a convolutional block followed by a Sigmoid activation function,

$$S^S = \text{Sigmoid}\big(\text{Conv}(V^S)\big). \tag{10}$$

Channel Attention Learning: We first process input feature map F_i by average pooling as:

$$X_i^S = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} f_{w,h}. \tag{11}$$

Then following SENet [18], the channel attention mask $C^S \in \mathbb{R}^{C\times1\times1}$ is generated by two cascaded FC layers,

$$C^S = \text{Sigmoid}\Big(\text{FC}_2(\text{FC}_1(X_i^S))\Big). \tag{12}$$

With the proposed SRA, our pyramid aggregation structure can fuse spatial discriminative cues, improving the representation capacity of aggregated features.

3.2.2 Temporal Reference Attention

Adjacent frames have strong temporal correlations, which can complement and enhance each other. However, most of existing works [24, 29] obtain the temporal feature by the average pooling, which may suffer huge interference when occlusion occurs. To this end, we propose a novel Temporal Reference Attention (TRA) to explore temporal relations.

As shown in Fig. 3, TRA has a similar structure with SRA. We also factorize A_{ij}^n into two low-dimension attention masks as:

$$A_t^T = S_t^T \circ C_t^T, \tag{13}$$

where $S_t^T \in \mathbb{R}^{1\times H\times W}$ and $C_t^T \in \mathbb{R}^{C\times1\times1}$ represent the spatial and channel attention mask, respectively.

Spatial Attention Learning: Given a pair of temporal adjacent feature maps F_t, F_{t+1}, we construct two graphs G_t, G_{t+1} with $N = W \times H$ nodes. Each node has a C-dimensional attribute vector $x_{t,i} \in \mathbb{R}^C$, where $i = 1, \cdots, N$. In addition to the embedding feature, structural relation is proved as an effective cue to learn the attention [54]. To make full use of relevant information between the adjacent frames, we compare a node in the current frame with all the nodes in its adjacent frames. Then we stack all bi-directional similarities to form the relation vector,

$$\begin{aligned} r_{i,j}^t &= \theta^T(x_{t,i})^T \phi^T(x_{t+1,j}), \\ r_{j,i}^{t+1} &= \theta^T(x_{t+1,j})^T \phi^T(x_{t,i}), \\ r_i^t &= \gamma^T\Big(r_{i,1}^t \cdot\cdot, r_{i,N}^t, r_{1,i}^{t+1} \cdots r_{N,i}^{t+1}\Big), \end{aligned} \tag{14}$$

where r_i^t is the relation vector of node i in graph G_t. θ^T, ϕ^T and γ^T are three embedding functions. Then the relation-value vector $v_{t,i}^T$ is constructed by stacking the relation vector $r_{t,i}^T$ and its corresponding embedding vector,

$$v_{t,i}^T = \Big[r_{t,i}^T, \beta(x_{t,i})\Big]. \tag{15}$$

Notably, the parameters of $\beta(x_{t,i})$ are the same as the ones mentioned in SRA. Finally, the spatial attention map $S_t^T \in \mathbb{R}^{1\times H\times W}$ can be obtained by

$$S_t^T = \text{Sigmoid}\Big(\text{Conv}(V_t^S)\Big), \tag{16}$$

where V_t^T is the relation reference matrix constructed by placing each $v_{t,i}^S$ to its corresponding location of the input feature map. For $t+1$ time step, attention maps of F_{t+1} can be obtained by just exchanging the place of t and $t+1$.

Channel Attention Learning: Two input feature maps are first sent into an embedding layer for computational efficiency. Then an average pooling layer is employed. Thus two feature vectors X_t^T and X_{t+1}^T can be achieved. Afterward, to take account of channel-wise influence between adjacent inputs, we concentrate X_t^T and X_{t+1}^T as $X_{t,t+1}^T$. Then, following SENet [18], the channel attention $C_t^T \in \mathbb{R}^{2C}$ is generated by two cascaded FC layers,

$$C_t^T = \text{Sigmoid}\Big(\text{FC}_2\big(\text{FC}_1(X_{t,t+1}^T)\big)\Big). \tag{17}$$

Different from SENet [18] and RGA-S [54] that process the single image, our TRA focuses on extracting the mutual information between the adjacent frames. Besides, it is reasonable that TRA, together with SRA, can extract more discriminative spatial-temporal information.

3.3. Loss Functions

To optimize our framework, we adopt the following objective function ($N = 3$),

$$L_{total} = \frac{1}{N} \sum_{n=1}^{N} [L_{cls}(X^n) + L_{tri}(X^n)], \tag{18}$$

where L_{cls} and L_{tri} are the classification loss and triplet loss [14] respectively. We choose the cross-entropy loss with label smoothing [40] as the classification loss to learn the identify-specific representation and avoid overfitting. We also employ the triplet loss [14] with batch hard mining to improve the ranking performance.

4. Experiments

4.1. Datasets and Protocols

MARS [55] is the largest video-based person ReID dataset, which contains 17,503 tracklets from 1261 identities and additional 3,248 tracklets of poor quality serving as distracters captured by 6 cameras. The videos of the MARS dataset are generated by the DPM [9] detector and GMMCP [7] tracker. The training set contains 625 identities and the testing set contains 636 identities.

DukeMTMC-VID [47] is another large-scale dataset with 4,832 tracklets and 1,812 identities. It is derived from the DukeMTMC [36] dataset. Among them, 702 identities with 2,196 tracklets are used for training and 3,338 tracklets of the rest 702 identities are for testing.

iLIDS-VID [44] contains 300 persons captured by two non-overlapping cameras constituting 600 image sequences. The length of each video sequence varies from 23 to 192 frames, with an average duration of 73 frames.

PRID-2011 [15] includes 385 and 749 identities from two non-overlapping cameras respectively with only the first 200 identities appear in both cameras. The video length of the PRID-2011 is varied from 5 to 675 frames.

Evaluation Protocols. We adopt the mean Average Precision (mAP) and the Cumulative Matching Characteristics (CMC) to evaluate the performance.

Table 1. Comparison of different components on MARS [55] and DukeMTMC-VID [47]. In the second column, **AS** means the **A**ggregating **S**tructure and **P**, **A**, **G** are the abbreviation of **P**yramid, **A**djacent and **G**lobal respectively. In the fourth column, **Res** denotes that the residual block adopted in the STAM.

Baseline	AS	STAM			MARS		DukeMTMC-VID		Speed	Params	GFLOPs
		SRA	TRA	Res	mAP	Rank-1	mAP	Rank-1			
✓	–	×	×	✓	84.7	88.8	96.3	95.9	129.29 Clip/s	26.24M	34.57
✓	**P**	×	×	✓	85.2	90.6	96.7	97.3	82.58 Clip/s	31.77M	35.58
✓	**A**	✓	✓	✓	85.1	90.2	96.5	96.8	80.00 Clip/s	32.37M	35.84
✓	**G**	✓	✓	✓	85.2	90.1	96.6	97.0	81.01 Clip/s	29.78M	35.84
✓	**P**	✓	×	✓	85.4	90.9	96.9	97.7	79.01 Clip/s	33.55M	36.27
✓	**P**	×	✓	✓	85.5	91.2	97.1	97.9	79.01 Clip/s	33.65M	36.27
✓	**P**	✓	✓	×	85.8	91.3	97.4	98.1	78.05 Clip/s	32.07M	35.86
✓	**P**	✓	✓	✓	**85.8**	**91.5**	**97.4**	**98.3**	78.04 Clip/s	35.42M	36.86

4.2. Implementation Details

This work implements our model based on Pytorch toolbox[1] on NVIDIA GTX 2080Ti GPUs (11GB memory). Following the RRS strategy [28], we sample 8 frames from the input video. And each frame is resized to 256×128 and augmented by random erasing and normalization. For the triplet loss, we randomly select 8 persons and sample 4 video clips for each individual. We employ the ResNet-50 [13] pre-trained on ImageNet [20] as our backbone network. Following the setting in [33], we set the last stride of ResNet-50 to 1 and remove the last spatial down-sample operation. During the training, we apply Adam [35] with weight decay 5×10^{-4} to update the parameters. We set the initial learning rate as 3.5×10^{-4} and follow the learning rate decay strategy as [33]. The model is trained for 500 epochs in total. During testing, we employ the cosine similarity for measuring the distance between query and gallery.

4.3. Ablation Study

4.3.1 Component Analysis

To verify the effect of each component, we conduct ablation experiments on MARS [55] and DukeMTMC-VID [47] datasets, shown in Tab. 1. Our baseline method only employs cross-entropy loss and triplet loss [14] on the average of the frame-level features and takes it as the video feature during testing. The results of the baseline method are shown in the first row of Tab. 1.

Effectiveness of Pyramid Aggregation Structure. To verify the effectiveness of our Pyramid Aggregation Structure (PAS), we first remove SRA and TRA from our full model for a fair comparison with the baseline method. Specifically, we adopt the pyramid structure but aggregate the frame-level features with a simple module which consists of an element-wise addition operation followed by a residual convolutional block. The results are shown in the second row of Tab. 1. Comparing with the baseline, our PAS improves the Rank-1 by 1.8% on MARS and 1.4% on DukeMTMC-VID. To further investigate the superiority of our method, We compare our full model with two simplified Aggregation Structures (AS), including Adjacent Reference Aggregation Structure (ARAS) and Global Reference Aggregation Structure (GRAS). ARAS extracts the local spatial-temporal information by referencing adjacent frames, and averages the refined feature to represent the whole sequence. While GRAS uses the average of all frame features as a reference. As shown in the 3-*th*, 4-*th*, and 7-*th* rows in Tab. 1, our PSTA outperforms ARAS and GRAS significantly by 1.3% and 1.4% in Rank-1 on MARS, as well as 1.5% and 1.3% on DukeMTMC-VID. We argue that such simplified aggregation structures lose the interaction between long-term and short-term relations, limiting their capability to mine the potential information in a video. But, our PSTA obtain the global dependence by progressively aggregating local information, which suppress the interference features and enhance the discriminative ones.

Effectiveness of Key Components of STAM. We evaluate the contribution of each component and report the results in Tab. 1. As shown in the 5-*th* and 6-*th* rows , we employ the SRA module and TRA module solely and respectively. Compared with the method only employing PAS, applying SRA and TRA respectively can further improve the Rank-1 score by 0.3% and 0.6% on MARS as well as 0.4% and 0.6% on DukeMTMC-VID. Finally, our PSTA improves the Rank-1 accuracy by 2.7% and 2.4%. Moreover, we remove the residual block applied in each layer of STAM to evaluate the influence of the convolution block. As shown in the 7-*th* row, the performance of removing the residual block is close to that of the whole network. It means that the improvement of our network is mainly from our pyramid temporal aggregation structure and well-designed spatial aggregation module instead of simply stacking layers.

Table 2. Performance of PSTA on MARS and DukeMTMC-VID under different number of STAM stages.

Stages	MARS		DukeMTMC-VID	
	mAP	rank-1	mAP	rank-1
0	84.7	88.8	96.5	96.9
1	85.2	90.0	96.7	97.0
2	85.6	91.0	97.0	97.6
3	**85.8**	**91.5**	**97.4**	**98.3**

Effect of Different Numbers of STAM. We investigate the influence of the numbers of STAM stages in Tab. 2. Note

[1] http://pytorch.org/

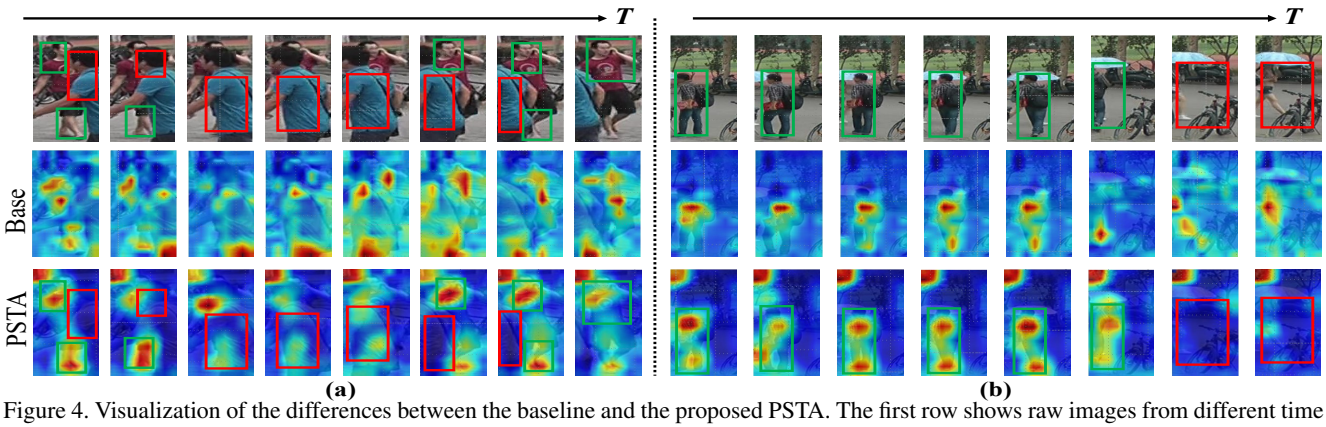


Figure 4. Visualization of the differences between the baseline and the proposed PSTA. The first row shows raw images from different time stamps. The second and the third rows are the channel activation maps of the baseline and PSTA respectively.

that when the number of STAM stages is set to 0, the structure of the model is the same as the baseline. As the number of STAM stages increases, there is a general improvement in performance. We argue that it is because multiple STAM stages can capture more comprehensive information from longer temporal dependence. Following the definition of STAM, the maximum number n of STAM stages should satisfy $T = 2^n$, where T is the length of the sequence. Thus, we set the sequence length and the number of STAM stages to 8 and 3 respectively.

Complexity Analysis. As shown in Tab. 1, comparing with the baseline, PSTA introduces additional $9M$ parameters and $2.29G$ computational complexity (FLOPs). The count of additional parameters and computational complexity is correlated with the complexity of the STAM.

4.3.2 Visualization Analysis

Visualization of Activation Maps. In Fig. 4, we visualize the channel activation maps of the baseline and the PSTA on MARS. As shown in Fig. 4 (a), compared to the baseline, it can be observed that the PSTA can further suppress the occlusion frames and enhance the discriminative features which are annotated by red and green bounding boxes. In Fig. 4 (b), we can find that, different from the baseline, PSTA focuses on the target person and avoids background information interference based on spatial and temporal reference attention. Furthermore, with the help of long-term information, PSTA can suppress the influence of irrelevant cues, which appear in the last two images in Fig. 4 (b).

Visualization of Retrieval Results. We visualize the retrieval results of a hard sample with different methods in Fig. 5 and conduct three experiments to demonstrate the effectiveness of the proposed PSTA. As can be observed, it is difficult for the baseline model to distinguish persons when the occluded frames appear. As shown in the second and the third column of Fig. 5, the visual ambiguity of the query is reduced in the top-1 retrieval result. However, ARAS and GRAS utilize simplified structures when aggregating frame-level features, which may lose some important local information (*e.g.* facial features). PSTA further considers the local and global temporal relation in these cases and employs the pyramid structure to extract more discriminative information. Besides, the retrieval results prove that the proposed PSTA indeed alleviates the problem of occlusion and captures the global long-term information.

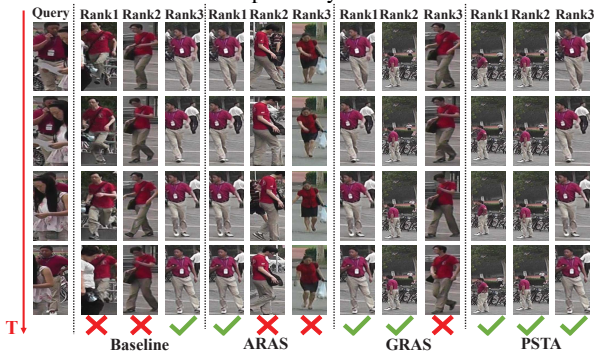
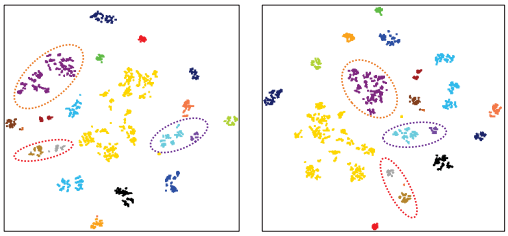


Figure 5. Visual examples on MARS dataset. Each example shows the top-3 retrieval image sequences by the baseline, ARAS, GRAS and, PSTA respectively. The true and wrong match are annotated by the green tick and red cross. **Best viewed in color.**



(a) Baseline (b) **PSTA**

Figure 6. Feature distribution of the baseline and PSTA visualized by t-SNE. We selected 20 pedestrians with a similar appearance from MARS. Each dot of different colors represents a different identity. We use three virtual coils of different colors to mark three distinct feature distributions.

Visualization of Feature Distribution. To further demonstrate the interpretability, we visualize the distribution of the final video-level feature extracted by the baseline method and our PSTA using t-SNE [41] in Fig. 6. Each point indicates the feature of a tracklet with 8 frames sampled from a video sequence. The identity of a point is indicated by its color. Compared with baseline (a), our model can better reduce intra-class distance and increase

inter-class distance. Especially for some points with fuzzy judgment in the baseline, our method can get better feature expression. Specifically, as shown in the orange circles of Fig. 6 (a) and Fig. 6 (b), the proposed PSTA framework can significantly reduce the intra-class variance. We infer that it is because PSTA could alleviate the interference of outlier step-by-step with the pyramid structure and thus reduce the intra-class distance. Besides comparing the distribution of red and blue circles, PSTA enlarge the inter-class distance.

4.4. Comparison with State-of-the-arts

In this section, we compare the proposed PSTA with other state-of-the-art methods on four video-based person ReID benchmarks: iLIDS-VID [44], PRID-2011 [15], MARS [55] and DukeMTMC-VID [47]. The results are reported in Tab. 3 and Tab. 4. Note that no post-processing techniques, such as re-ranking [56] or multi-query [55] are employed in these experiments.

Table 3. Performance (%) comparison of our method with state-of-the-art methods on MARS [55] and DukeMTMC-VID [47].

Model	MARS		DukeMTMC-VID	
	mAP	Rank-1	mAP	Rank-1
CNN+XQDA[55]	47.6	65.3	–	–
STIM [31]	72.7	84.4	–	–
M3D [23]	74.1	84.4	–	–
STA [10]	80.8	86.3	94.9	96.2
AMEM [25]	79.3	86.7	–	–
COSAM [38]	79.9	84.9	94.1	95.4
GLTR [22]	78.5	87.0	93.7	96.3
RTF [21]	85.2	87.1	–	–
FGRA [5]	81.2	87.3	–	–
VRSTC [17]	82.3	88.5	93.5	95.0
STE-NVAN [28]	81.2	88.9	–	–
TCLNet [16]	85.1	89.8	96.2	96.9
STGCN [50]	83.7	89.9	95.7	97.2
AFA [4]	82.9	90.2	95.4	97.2
MGH [49]	85.8	90.0	–	–
MG-RAFA [53]	**85.9**	88.8	–	–
AP3D [12]	85.6	90.7	96.1	97.2
PSTA	85.8	**91.5**	**97.4**	**98.3**

On MARS and DukeMTMC-VID, our PSTA method achieves competitive results compared with other state-of-the-art methods. More remarkably, the proposed PSTA achieves 91.5% in rank-1 accuracy on MARS, outperforming most of the published methods, *e.g.*, AP3D [12] and MG-RAFA [53], by 0.8% and 1.7% in Rank-1 accuracy. Note that AP3D [12] uses 3D CNN to learn the temporal cues, consequently requiring higher computational complexity. MG-RAFA [53] obtains a strong performance on mAP but with a significant gap with our method in Rank-1 accuracy. This may be because MG-RAFA adopts multi-granularity references which would capture the semantics of different levels and achieve high performance in mAP. However, the global reference used in [53] would introduce huge interference when long-term occlusion occurs, which is of common cases in video-based person ReID. The proposed PSTA captures the global dependence by progressively aggregating the local features while maintaining the information from short-term aggregation. Thus, it is more robust to long-term occlusions. In addition, GLTR [22] adopts the pyramid structure along with the temporal dimension by dilated convolutions. Such a way may lose the message passing from short-term relation to long-term dependence. Thus the accuracy is not so satisfactory.

Table 4. Performance (%) comparison of our method with state-of-the-art methods on iLIDS-VID [44] and PRID-2011 [15].

Model	iLIDS-VID		PRID-2011	
	Rank-1	Rank-5	Rank-1	Rank-5
CNN+XQDA[55]	53.0	84.4	77.3	93.5
STIM [31]	84.3	96.8	92.7	98.8
M3D [23]	74.0	94.3	94.4	**100**
AMEM [25]	87.2	97.7	93.3	98.7
COSAM [38]	79.6	95.3	–	–
GLTR [22]	86.0	98.0	95.5	**100**
Jiang *et al.* [21]	87.7	–	95.8	–
FGRA [5]	88.0	96.7	95.5	**100**
VRSTC [17]	83.4	–	–	–
TCLNet [16]	86.6	–	–	–
AFA [4]	88.5	96.8	–	–
MGH [49]	85.6	97.1	94.8	99.3
MG-RAFA [53]	88.6	98.0	**95.9**	99.7
AP3D [12]	88.7	–	–	–
PSTA	**91.5**	**98.1**	95.6	98.9

In terms of iLIDS-VID and PRID-2011 datasets, we only report the cumulative accuracy because the datasets have only one correct match in the gallery set. Our PSTA method achieves competitive performance compared to other methods with 91.5% and 95.6% Rank-1 accuracy on iLIDS-VID and PRID-2011 datasets respectively. Specifically, PSTA outperforms the best Rank-1 accuracy of published methods on iLIDS-VID by 2.94%.

5. Conclusions

This work presents a novel framework (PSTA) to fuse frame-level features progressively. Benefit from the pyramid structure, long-term dependence can be established without losing the useful local information. Furthermore, a Spatial-Temporal Aggregation Module is proposed. It contains two key components, a Spatial Reference Attention to generate attention maps with intra-frame relations and a Temporal Reference Attention to suppress the irrelevant feature and enhance the discriminative feature with inter-frame relations. Finally, extensive experiments demonstrate the superiority of our PSTA on several benchmarks.

Acknowledgments This work is in part supported by the Postgraduate Research & Practical Innovation Program of Jiangsu Province (No. KYCX20_3083), the National Science Foundation of China (No. 62006098) and the Fundamental Research Funds for the Central Universities (No. DUT20RC(3)083).

References

[1] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, pages 6299–6308, 2017.

[2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed High-Order Attention Network for Person Re-Identification. In *ICCV*, pages 371–381, 2019.

[3] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Learning Recurrent 3D Attention for Video-Based Person Re-Identification. *TIP*, 29:6963–6976, 2020.

[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.

[5] 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.

[6] Ju Dai, Pingping Zhang, Dong Wang, Huchuan Lu, and Hongyu Wang. Video Person Re-Identification by Temporal Residual Learning. *TIP*, 28(3):1366–1377, 2019.

[7] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. GMMCP Tracker: Globally Optimal Generalized Maximum Multi Clique Problem for Multiple Object Tracking. In *CVPR*, pages 4091–4099, 2015.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[9] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *CVPR*, pages 1–8, 2008.

[10] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. STA: Spatial-Temporal Attention for Large-Scale Video-Based Person Re-Identification. In *AAAI*, pages 8287–8294, 2019.

[11] Jiyang Gao and Ram Nevatia. Revisiting Temporal Modeling for Video-Based Person ReID. In *BMVC*, 2018.

[12] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-Preserving 3D Convolution for Video-Based Person Re-Identification. In *ECCV*, pages 228–243, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.077737*, 2017.

[15] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In *SCIA*, pages 91–102, 2011.

[16] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal Complementary Learning for Video Person Re-Identification. In *ECCV*, pages 388–405, 2020.

[17] 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.

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *CVPR*, pages 7132–7141, 2018.

[19] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, pages 448–456, 2015.

[20] Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[21] 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.

[22] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-Local Temporal Representations for Video Person Re-Identification. In *CVPR*, pages 3958–3967, 2019.

[23] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-Scale 3D Convolution Network for Video-Based Person Re-Identification. In *AAAI*, pages 8618–8625, 2019.

[24] 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.

[25] Shuzhao Li, Huimin Yu, and Haoji Hu. Appearance and Motion Enhancement for Video-Based Person Re-Identification. In *AAAI*, pages 11394–11401, 2020.

[26] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-Identification. In *CVPR*, pages 2285–2294, 2018.

[27] Xingyu Liao, Lingxiao He, Zhouwang Yang, and Chi Zhang. Video-Based Person Re-identification via 3D Convolutional Networks and Non-local Attention. In *ACCV*, pages 620–634, 2018.

[28] Chih-Ting Liu, Chih-Wei Wu, Yu-chiang Frank Wang, and Shao-Yi Chien. Spatially and Temporally Efficient Non-Local Attention Network for Video-Based Person Re-Identification. In *BMVC*, 2019.

[29] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based Person Re-Identification with Accumulative Motion Context. *TCSVT*, 28(10):2788–2802, 2018.

[30] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, Xuesheng Qian, and Xiaoyu Yang. A video is worth three views: Trigeminal transformers for video-based person re-identification. *arXiv preprint arXiv:2104.01745*, 2021.

[31] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and Temporal Mutual Promotion for Video-Based Person Re-Identification. In *AAAI*, pages 8786–8793, 2019.

[32] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, pages 2448–2457, 2017.

[33] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *CVPR workshop*, 2019.

[34] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent Convolutional Network for Video-Based Person Re-Identification. In *CVPR*, pages 1325–1334, 2016.

[35] Diederik P.Kingma and Jimmyy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2014.

[36] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data set for Multi-Target, Multi-Camera Tracking. In *ECCV workshop*, pages 17–35, 2016.

[37] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-Identification. In *CVPR*, pages 5363–5372, 2018.

[38] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-Segmentation Inspired Attention Networks for Video-Based Person Re-Identification. In *ICCV*, pages 562–572, 2019.

[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.

[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.

[41] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *MLR*, 9(86):2579–2605, 2008.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NIPS*, pages 5998–6008, 2017.

[43] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-Temporal Person Re-Identification. In *AAAI*, pages 8933–8940, 2019.

[44] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person Re-Identification by Video Ranking. In *ECCV*, pages 688–703, 2014.

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-Local Neural Networks. In *CVPR*, pages 7794–7803, 2018.

[46] Yiming Wu, Omar EI Farouk Bourahla, Xi Li, Fei Wu, and Qi Tian. Adaptive Graph Representation Learning for Video Person Re-Identification. *TIP*, 29:8821–8830, 2020.

[47] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the Unknown Gradually : One-Shot Video-Based Person Re-Identification by Stepwise Learning. In *CVPR*, pages 5177–5186, 2018.

[48] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly Attentive Spatial-Temporal Pooling Networks for Video-Based Person Re-Identification. In *ICCV*, pages 4733–4742, 2017.

[49] 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.

[50] 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.

[51] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, 2021.

[52] Le Zhang, Shi Zenglin, Joey Tianyi Zhou, Ming-Ming Cheng, Yun Liu, Jia-Wang Bian, Zeng Zeng, and Chunhua Shen. Ordered or Orderless: A Revisit for Video based Person Re-Identification. *TPAMI*, 43(4):1460–1466, 2021.

[53] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-Based Person Re-Identification. In *CVPR*, pages 10407–10416, 2020.

[54] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-Aware Global Attention for Person Re-Identification. In *CVPR*, pages 3186–3195, 2020.

[55] 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, 2016.

[56] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-Ranking Person Re-Identification With k-Reciprocal Encoding. In *CVPR*, pages 1318–1327, 2017.

[57] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-Identification. In *CVPR*, pages 4747–4756, 2017.