







# Channel Recurrent Attention Networks for Video Pedestrian Retrieval

Pengfei Fang<sup>1,2</sup>(✉) , Pan Ji<sup>3</sup> , Jieming Zhou<sup>1,2</sup> , Lars Petersson<sup>2</sup> ,  
and Mehrtash Harandi<sup>4</sup> 

<sup>1</sup> The Australian National University, Canberra, Australia

[Pengfei.Fang@anu.edu.au](mailto: Pengfei.Fang@anu.edu.au)

<sup>2</sup> DATA61-CSIRO, Sydney, Australia

<sup>3</sup> OPPO US Research Center, Melbourne, Australia

<sup>4</sup> Monash University, Melbourne, Australia

**Abstract.** Full attention, which generates an attention value per element of the input feature maps, has been successfully demonstrated to be beneficial in visual tasks. In this work, we propose a fully attentional network, termed *channel recurrent attention network*, for the task of video pedestrian retrieval. The main attention unit, *channel recurrent attention*, identifies attention maps at the frame level by jointly leveraging spatial and channel patterns via a recurrent neural network. This channel recurrent attention is designed to build a global receptive field by recurrently receiving and learning the spatial vectors. Then, a *set aggregation* cell is employed to generate a compact video representation. Empirical experimental results demonstrate the superior performance of the proposed deep network, outperforming current state-of-the-art results across standard video person retrieval benchmarks, and a thorough ablation study shows the effectiveness of the proposed units.

**Keywords:** Full attention · Pedestrian retrieval · Channel recurrent attention · Global receptive field · Set aggregation

## 1 Introduction

This work proposes *Channel Recurrent Attention Networks* for the purpose of pedestrian retrieval<sup>1</sup>, in challenging video data.

Pedestrian retrieval, or person re-identification (re-ID), a core task when tracking people across camera networks [1], attempts to retrieve all correct

<sup>1</sup> For the remainder of this paper, we shall use the terms “pedestrian retrieval” and “person re-identification” interchangeably.

P. Ji—Work done while at NEC Laboratories America.

**Electronic supplementary material** The online version of this chapter ([https://doi.org/10.1007/978-3-030-69544-6\\_26](https://doi.org/10.1007/978-3-030-69544-6_26)) contains supplementary material, which is available to authorized users.

matches of a person from an existing database, given a target query. There are many challenges to this task, with a majority stemming from a poor quality or large variation of the captured images. This often leads to difficulties in building a discriminative representation, which in turn results in a retrieval system to mismatch its queries. Video-, as opposed to single image-, person re-ID offers the possibility of a richer and more robust representation as temporal cues can be utilized to obtain a compact, discriminative and robust video representation for the re-ID task. In many practical situations, the retrieval performance suffers from spatial misalignment [2–4], caused by the movement of body parts, which affects the retrieval machine negatively. Focusing on this issue, many efforts have been made to develop visual attention mechanisms [3, 5–8], which makes the network attend to the discriminative areas within person bounding boxes, relaxing the constraints stemming from spatial nuances.

Attention mechanisms have been demonstrated to be successful in various visual tasks, such as image classification [9, 10], object detection [11], scene segmentation [12, 13] to name just a few. Generally speaking, attention mechanisms can be grouped into channel attention [9], spatial attention [14], and full attention [5], according to the dimensions of the generated attention maps. The channel attention usually summarizes the global spatial representation of the input feature maps, and learns a channel pattern that re-weights each slice of the feature maps. In contrast, the spatial attention learns the spatial relationships within the input feature maps and re-weights each spatial location of the feature maps. Lastly, full attention not only learns the channel patterns, but also preserves spatial information in the feature maps, which significantly improves the representation learning [15].

Various types of full attention mechanisms have been studied extensively for the task of pedestrian retrieval [5–7]. In [5], the fully attentional block re-calibrates the channel patterns by a non-linear transformation. Thereafter, higher order channel patterns are exploited to attend to the channel features [6, 7]. However, the aforementioned attention fails to build *long-range* spatial relationships due to the use of a  $1 \times 1$  convolution. The work in [3] learns spatial interactions via a convolutional layer with a larger kernel size ( $3 \times 3$ ), but the attention module therein still only has a small spatial receptive field. In visual attention, we want the network to have the capacity to view the feature maps globally and decide what to focus on for further processing [11]. A global view can be achieved by applying fully connected (FC) layers, which, unfortunately, introduces a huge number of learnable parameters if implemented naively.

In this work, we propose a full attention mechanism, termed *channel recurrent attention*, to boost the video pedestrian retrieval performance. The channel recurrent attention module aims at creating a global view of the input feature maps. Here, the channel recurrent attention module benefits from the recurrent operation and the FC layer in the recurrent neural network. We feed the vectorized spatial map to the Long Short Term Memory (LSTM) sequentially, such that the recurrent operation of the LSTM captures channel patterns while the FC layer in the LSTM has a global receptive field of each spatial slice. To handle

video data, we continue to develop a *set aggregation* cell, which aggregates the frame features into a discriminative clip representation. In the set aggregation cell, we re-weight each element of the corresponding frame features, in order to selectively emphasize useful features and suppress less informative features, with the aid of the associated clip features. The **contributions** of this work include:

- The proposal of a novel channel recurrent attention module to jointly learn spatial and channel patterns of each frame feature map, capturing the global view of the feature maps. To the best of the authors’ knowledge, this is the first attempt to consider the global spatial and channel information of feature maps in a full attention design for video person re-ID.
- The development of a simple yet effective set aggregation cell, which aggregates a set of frame features into a discriminative clip representation.
- State-of-the-art performance across standard video re-ID benchmarks by the proposed network. The generalization of the attention module is also verified by the competitive performance on the single image re-ID task.

## 2 Related Work

This section summarizes the related work of pedestrian retrieval and relevant attention mechanisms.

Several approaches have been investigated to improve the state-of-the-art retrieval performance for both single image and video person re-ID [1, 16]. Focusing on deep neural networks [17], metric learning and representation learning are the two dominating approaches in modern person re-ID solutions [1]. In [18], the similarity of input pairs is calculated by a siamese architecture [19]. The improved deep siamese network learns an image-difference metric space by computing the cross-input relationships [20].

In representation learning, the pedestrian is represented by the concatenation of multi-level features along the deep network [21] or by combining the person appearance and body part features [3]. Beyond single image-based person re-ID methods, efficient temporal modeling [22] is further required when working with video clips. This is challenging as there is a need to create a compact video representation for each identity. McLaughlin *et al.* proposed average and max temporal pooling for aggregating frame features and each frame feature is the output of a recurrent neural network [23].

Recent work has shown that person re-ID benefits significantly from attention mechanisms highlighting the discriminative areas inside the person bounding boxes when learning an embedding space [3, 5–7, 24, 25]. In [24, 25], the spatial attention mask is designed to attend one target feature map or various feature maps along the deep network. In [5], a fully attentional block is developed to re-calibrate the channel features. Second or higher order statistical information is also employed in full attention frameworks [6, 7]. The full attention shape map is also generated in the harmonious attention module [3], by integrating channel attention and spatial attention. The aforementioned attention mechanism either fails to build spatial-wise relationships, or receives a limited spatial receptive

field. Unlike the above methodology of full attention, we intend to develop an attention mechanism which preserves the advantage of the common full attention, while also perceiving a global spatial receptive field of the feature maps.

### 3 Method

This section details the proposed deep network in a top-down fashion: starting with the problem formulation of the application, followed by the network architecture and the main attention module in the network, namely, the channel recurrent attention module. Thereafter, we also introduce a set aggregation cell, to encode a compact clip representation.

**Notation.** We use  $\mathbb{R}^n$ ,  $\mathbb{R}^{h \times w}$ ,  $\mathbb{R}^{c \times h \times w}$  and  $\mathbb{R}^{t \times c \times h \times w}$  to denote the  $n$ -dimensional Euclidean space, the real matrix space (of size  $h \times w$ ), and the image and video spaces, respectively. A matrix or vector transpose is denoted by the superscript  $\top$ . The symbol  $\odot$  and  $\oplus$ , represent the Hadamard product (*i.e.*, element-wise multiplication) and element-wise summation.  $\sigma : \mathbb{R} \rightarrow [0, 1]$  is the sigmoid function.  $\text{BN} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $\text{BN}(\mathbf{x}) := \gamma \frac{\mathbf{x} - \mathbb{E}[\mathbf{x}]}{\sqrt{\text{Var}[\mathbf{x}]}} + \beta$  and  $\text{ReLU} : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ ,  $\text{ReLU}(x) := \max(0, x)$  refer to batch normalization and rectified linear unit.  $\phi$ ,  $\varphi$ ,  $\varpi$  and  $\psi$  are used to represent embedding functions (*e.g.*, linear transformations, or self-gating layer).

#### 3.1 Problem Formulation

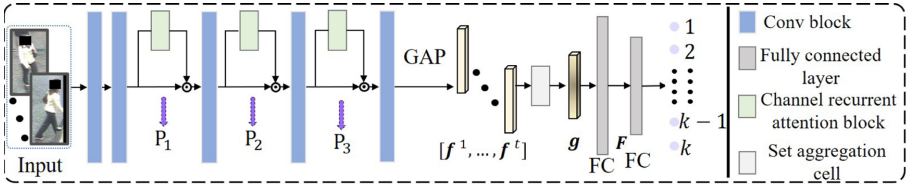
Let a fourth-order tensor,  $\mathbf{T}_i = [T_i^1, T_i^2, \dots, T_i^N] \in \mathbb{R}^{N \times C \times H \times W}$ , denote the  $i$ -th video sequence of a pedestrian, where  $N$ ,  $C$ ,  $H$ , and  $W$  are the number of frames, channels, height and width, respectively. Each video sequence  $\mathbf{T}_i$  is labeled by its identity, denoted by  $y_i \in \{1, \dots, k\}$ . The training set with  $M$  video sequences is described by  $\mathbb{T} = \{\mathbf{T}_i, y_i\}_{i=1}^M$ . The video person re-ID model,  $f_\theta : \mathcal{T} \rightarrow \mathcal{F}$ , describes a non-linear embedding from the video space,  $\mathcal{T}$ , to an embedding space,  $\mathcal{F}$ , in which the intra-class/person distance is minimized and the inter-class/person distance is maximized. The target of training a deep neural network is to learn a set of parameters,  $\theta^*$ , with minimum loss value (*e.g.*,  $\mathcal{L}$ ), satisfying:  $\theta^* = \arg \min_\theta \sum_{i=1}^M \mathcal{L}(f_\theta(\mathbf{T}_i), y_i)$ . In the training stage, we randomly sample batches of video clips, where each video clip has only  $t$  frames (randomly chosen). Such frames are order-less and hence, we are interested in set-matching for video re-ID.

#### 3.2 Overview

We begin by providing a sketch of our design first. In video person re-ID, one would ideally like to make use of a deep network to extract the features of the frames and fuse them into a compact and discriminative clip-level representation. In the lower layers of our design, we have five convolutional blocks along with channel recurrent attention modules at positions  $P_1$ ,  $P_2$  and  $P_3$  (see Fig. 1). Once

the deep network extracts a set of frame features (*i.e.*,  $[f^1, \dots, f^t]$  in Fig. 1), a set aggregation cell is utilised to fuse frame features into a compact clip-level feature representation (*i.e.*,  $\mathbf{g}$ ). The final clip representation is  $\mathbf{F} = \text{ReLU}(\text{BN}(\mathbf{W}_1^\top \mathbf{g}))$ , followed by another FC layer to perform identity prediction (*i.e.*,  $p = \mathbf{W}_2^\top \mathbf{F}$ ), where  $\mathbf{W}_1, \mathbf{W}_2$  are the learnable parameters in the FC layers. We note that the output of the middle convolutional layers captures rich spatial and channel information [6, 11], such that the attention modules can make better use of this available information.

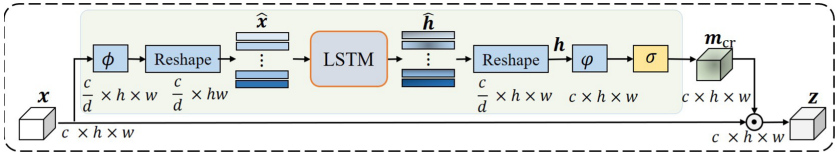
The network training benefits from multi-task learning, which formulates the network training as several sub-tasks. Our work follows [22], and trains the network using a triplet loss and a cross-entropy loss. The details of the loss functions are described in the supplementary material.



**Fig. 1.** The proposed deep neural network with channel recurrent attention modules and a set aggregation cell.

### 3.3 Channel Recurrent Attention

We propose the channel recurrent attention module (see Fig. 2), which learns the spatial and channel patterns globally in a collaborative manner with the assistance of an LSTM, over the feature maps of each frame. To be specific, we model the input feature maps as a sequence of spatial feature vectors, and feed it to an LSTM to capture global channel patterns by its recurrent operation. In our design, the hidden layer (*e.g.*, FC) of the LSTM unit, can be understood as having a global receptive field, acting on each spatial vector while sharing weights with other spatial vectors, addressing the limitation of a small receptive field in CNNs. In Sect. 4, our claim is empirically evaluated in an ablation study.



**Fig. 2.** The architecture of the proposed channel recurrent attention module.

Let  $\mathbf{x} \in \mathbb{R}^{c \times h \times w}$  be the input of the channel recurrent attention module. In our implementation, we project  $\mathbf{x}$  to  $\phi(\mathbf{x})$ , reducing the channel dimension by a ratio of  $1/d$ , and reshape the embedded tensor  $\phi(\mathbf{x})$  to a matrix  $\hat{\mathbf{x}} = [\hat{x}_1, \dots, \hat{x}_{\frac{c}{d}}]^\top \in \mathbb{R}^{\frac{c}{d} \times hw}$ , where a row of  $\hat{\mathbf{x}}$  (e.g.,  $\hat{x}_i \in \mathbb{R}^{hw}, i = 1, \dots, \frac{c}{d}$ ) denotes the spatial vector of a slice. The effect of the ratio  $1/d$  is studied in Sect. 4.2. A sequence of spatial vectors is then fed to an LSTM unit and the LSTM generates a sequence of hidden states, in matrix form:

$$\hat{\mathbf{h}} = \text{LSTM}(\hat{\mathbf{x}}) = [\hat{h}_1, \dots, \hat{h}_{\frac{c}{d}}]^\top, \tag{1}$$

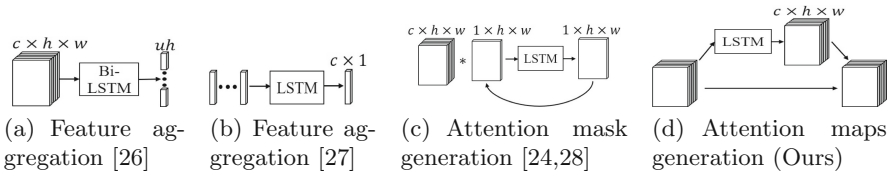
where  $\hat{h}_i \in \mathbb{R}^{hw}, i = 1, \dots, c/d$  is a sequence of hidden states and  $\text{LSTM}(\cdot)$  represents the recurrent operation in an LSTM. The insight is illustrated by the unrolled LSTM, shown in Fig. 5(a).  $\hat{\mathbf{h}}$  is further reshaped to the same size as the input tensor  $\phi(\mathbf{x})$  (i.e.,  $\mathbf{h} = \text{Reshape}(\hat{\mathbf{h}}), \mathbf{h} \in \mathbb{R}^{\frac{c}{d} \times h \times w}$ ). The final attention value is obtained by normalizing the embedded  $\mathbf{h}$ , written as:

$$\mathbf{m}_{\text{cr}} = \sigma(\varphi(\mathbf{h})). \tag{2}$$

Here,  $\varphi(\mathbf{h}), \mathbf{m}_{\text{cr}} \in \mathbb{R}^{c \times h \times w}$ . This normalized tensor acts as a full attention map and re-weights the elements of the associated frame feature map (see Fig. 2), by element-wise multiplication:

$$\mathbf{z} = \mathbf{m}_{\text{cr}} \odot \mathbf{x}. \tag{3}$$

*Remark 1.* There are several studies that use LSTMs to aggregate features [26, 27] (see Fig. 3(a) and 3(b)), or generate attention masks [24, 28] (see Fig. 3(c)). Our channel recurrent attention module (see Fig. 3(d)) is significantly different from existing works as shown in Fig. 3. The designs in [26] and [27] employ an LSTM to aggregate features either from input feature maps [26], or a sequence of frame features in a video [27]. In [24, 28], an attention value for each spatial position of the feature maps (i.e., spatial attention) is constructed recursively, while ignoring the relation in the channel dimension. In contrast, our channel recurrent attention generates an attention value per element of the feature maps (i.e., full attention), thereby enabling the ability to learn richer spatial and channel features.



**Fig. 3.** Schematic comparison of our attention mechanism and existing LSTM-based works. In (c), the notation  $*$  denotes a weighted sum operation.

### 3.4 Set Aggregation

To encode a compact clip representation, we further develop a set aggregation cell to fuse the per frame features (see Fig. 4 for a block diagram). The set aggregation cell highlights the frame feature, with the aid of the clip feature, firstly, and then aggregates them by average pooling.

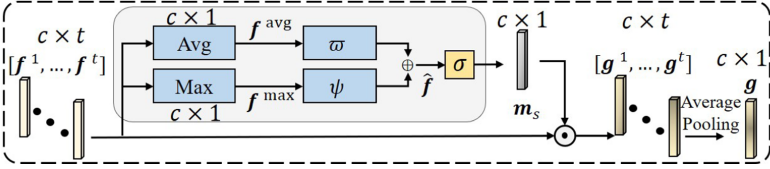


Fig. 4. The architecture of the proposed set aggregation cell.

Let  $[\mathbf{f}^1, \dots, \mathbf{f}^t]$ ,  $\mathbf{f}^j \in \mathbb{R}^c$  be a set of frame feature vectors, encoded by a deep network (see Fig. 1). The set aggregation cell first re-weights the frame features. In our implementation, we combine average pooling and max pooling to aggregate frame features. This is due to the fact that both pooling schemes encode different statistical information and their combination is expected to increase the representation capacity. More specifically, each element in  $\mathbf{f}^{\text{avg}}$  and  $\mathbf{f}^{\text{max}}$  are defined as  $f_i^{\text{avg}} = \text{avg}(f_i^1, \dots, f_i^t) = \frac{1}{t} \sum_{j=1}^t (f_i^j)$  and  $f_i^{\text{max}} = \max(f_i^1, \dots, f_i^t)$ , respectively. Each aggregation is followed by self-gating layers (*i.e.*,  $\varpi$  and  $\psi$  in Fig. 4) to generate per-element modulation weights, and fused by element-wise summation as:

$$\hat{\mathbf{f}} = \varpi(\mathbf{f}^{\text{avg}}) \oplus \psi(\mathbf{f}^{\text{max}}). \quad (4)$$

This is then followed by normalizing the fused weights to produce the final mask (*e.g.*,  $\mathbf{m}_s = \sigma(\hat{\mathbf{f}})$ ) which is applied as follows:

$$\mathbf{g}^j = \mathbf{m}_s \odot \mathbf{f}^j, \quad j = 1, \dots, t. \quad (5)$$

Finally, we use average pooling to obtain the clip feature,  $\mathbf{g} = 1/t \sum_{j=1}^t \mathbf{g}^j$ . We note that in our network the parameters in the two self-gating layers are not shared. This is to increase the diversity of features which is beneficial, and we evaluate it in Sect. 4.

*Remark 2.* The set aggregation cell is inspired by the Squeeze-and-Excitation (SE) block [9], in the sense that frame features will be emphasized under the context of the global clip-level features, but with a number of simple yet important differences: **(i)** The SE receives a feature map as input, while the input of our set aggregation is a set of frame features. **(ii)** The SE only uses global average pooling to encode the global feature of the feature maps, while the set aggregation employs both average and max pooling to encode hybrid clip features, exploiting more diverse information present in the frame features.

### 3.5 Implementation Details

**Network Architecture and Training.** We implemented our approach in the PyTorch [29] deep learning framework. We chose ResNet-50 [30] as the backbone network, pre-trained on ImageNet [31]. In a video clip with  $t$  frames, each frame-level feature map, produced by the last convolutional layer, is squeezed to a feature vector  $\mathbf{f}^j \in \mathbb{R}^{2048}$ ,  $j = 1, \dots, t$  by global average pooling (GAP). Subsequently, the set aggregation cell fuses the frame features to a compact clip feature vector  $\mathbf{g}$ . Following  $\mathbf{g}$ , the final clip-level person representation  $\mathbf{F}$  is embedded by a fully connected (FC) layer with the dimension 1024. Thereafter, another FC layer is added for the purpose of final classification during training. In the channel recurrent attention module, the ratio  $d$  is set to 16 for the PRID-2011 and iLIDS-VID datasets, and 8 for the MARS and DukeMTMC-VideoReID datasets, and the LSTM unit has one hidden layer. In the set aggregation cell, the self-gating layer is a bottleneck network to reduce the number of parameters, the dimension of the hidden vector is  $2048/r$ , and we choose  $r = 16$  as in [9], across all datasets. The ReLU and batch normalization are applied to each embedding layer and self-gating layer. The details of the datasets is described in Sect. 4.1.

We use the Adam [32] optimizer with default momentum. The initial learning rate is set to  $3 \times 10^{-4}$  for PRID-2011 and iLIDS0-VID, and  $4 \times 10^{-4}$  for MARS and DukeMTMC-VideoReID. The mini-batch size is set to 16 for the PRID-2011 and iLIDS-VID datasets and 32 for the MARS and DukeMTMC-VideoReID datasets, respectively. In a mini-batch, both  $P$  and  $K$  are set to 4 for the PRID-2011 and iLIDS-VID, whereas  $P = 8$ ,  $K = 4$  for the MARS and DukeMTMC-VideoReID. The margin in the triplet loss, *i.e.*,  $\xi$ , is set to 0.3 for all datasets. The spatial size of the input frame is fixed to  $256 \times 128$ . Following [22],  $t$  is chosen as 4 in all experiments and 4 frames are *randomly* sampled in each video clip [22, 33]. Our training images are randomly flipped in the horizontal direction, followed by random erasing (RE) [34]. We train the network for 800 epochs. The learning rate decay is set to 0.1, applied at the 200-th, 400-th epoch for the PRID-2011 and iLIDS-VID, and the 100-th, 200-th, 500-th epoch for the MARS and DukeMTMC-VideoReID, respectively. Moreover, it is worth noting that we do not apply re-ranking to boost the ranking result in the testing phase.

## 4 Experiment on Video Pedestrian Retrieval

### 4.1 Datasets and Evaluation Protocol

In this section, we perform experiments on four standard video benchmark datasets, *i.e.*, PRID-2011 [35], iLIDS-VID [36], MARS [37] and DukeMTMC-VideoReID [38] to verify the effectiveness of the proposed attentional network. **PRID-2011** has 400 video sequences, showing 200 different people where each person has 2 video sequences, captured by two separate cameras. The person bounding box is manually labeled. **iLIDS-VID** contains 600 image sequences of 300 pedestrians, captured by two non-overlapping cameras in an airport. Each



of the training and test sets has 150 person identities. In this dataset, the target person is heavily occluded by other pedestrians or objects (*e.g.*, baggage). **MARS** is one of the largest video person re-ID datasets which contains 1,261 different identities and 20,715 video sequences captured by 6 separate cameras. The video sequences are generated by the GMMCP tracker [39], and for each frame, the bounding box is detected by DPM [40]. The dataset is split into training and testing sets that contain 631 and 630 person identities, respectively. **DukeMTMC-VideoReID** is another large video person re-ID dataset. This dataset contains 702 pedestrians for training, 702 pedestrians for testing as well as 408 pedestrians as distractors. The training set and testing set has 2,196 video sequences and 2,636 video sequences, respectively. The person bounding boxes are annotated manually.

Following existing works, we use both the cumulative matching characteristic (CMC) curve and mean average precision (mAP) to evaluate the performance of the trained re-ID system.

## 4.2 Ablation Study

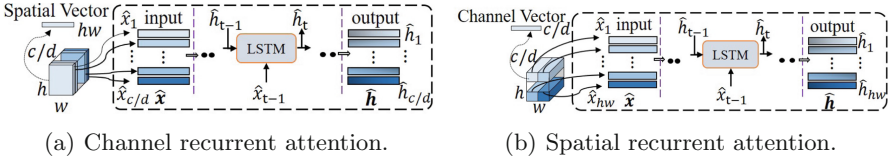
This section demonstrates the effectiveness of the proposed blocks and the selection of appropriate hyper parameters via a thorough battery of experiments.

**Effect of Channel Recurrent Attention.** Here, we evaluate the effectiveness of the proposed channel recurrent attention, and verify our claim that our channel recurrent attention is able to capture more structure information as we sequentially feed the spatial vector to the LSTM. To show the design is reasonable, we compare our channel recurrent attention with two variations, namely, the spatial recurrent attention and the conv attention.

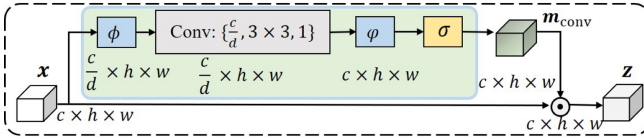
In the spatial recurrent attention, the LSTM receives a sequence of channel features from feature maps as input, with the recurrent operator along the spatial domain. In more detail, in channel recurrent attention (see Fig. 2), the input is a sequence of spatial vectors, (*e.g.*,  $\hat{\mathbf{x}} = [\hat{x}_1, \dots, \hat{x}_{\frac{c}{a}}]^\top \in \mathbb{R}^{\frac{c}{a} \times hw}$ ). In the spatial recurrent attention, the input is a sequence of channel vectors, (*e.g.*,  $\hat{\mathbf{x}} = [\hat{x}_1, \dots, \hat{x}_{hw}]^\top \in \mathbb{R}^{hw \times \frac{c}{a}}$ ). Though the recurrent operation along the spatial domain is also able to learn the pattern spatially, the spatial recurrent attention lacks explicit modeling in the spatial domain. Figure 5 shows the schematic difference between channel recurrent attention (see Fig. 5(a)) and spatial recurrent attention (see Fig. 5(b)).

In addition, to verify the necessity of a global receptive field in our channel recurrent attention, we further replace the LSTM with a convolutional layer with a similar parameter size, which is called a conv attention. The architecture of the conv attention is shown in Fig. 6. In the Conv block, the kernel size is  $3 \times 3$  and the sliding step is 1, and it produces a tensor with the shape of  $\frac{c}{a} \times h \times w$ . The generated attention mask can be formulated as  $\mathbf{m}_{\text{conv}} = \sigma\left(\varphi(\text{Conv}(\phi(\mathbf{x})))\right)$ , where  $\text{Conv}(\cdot)$  indicates the convolutional operation.

Table 1 compares the effectiveness of three attention variations. It is shown that our channel recurrent attention has a superior performance over the other



**Fig. 5.** Schematic comparison between channel recurrent attention and spatial recurrent attention.



**Fig. 6.** The architecture of the proposed conv attention module.

**Table 1.** Comparison of three attention variations across four datasets. CRA: Channel Recurrent Attention; SRA: Spatial Recurrent Attention; CA: Conv Attention.

Method		PRID-2011		iLIDS-VID		MARS		DukeMTMC-VideoReID	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
(i)	No Attention	85.4	91.0	80.0	87.1	82.3	76.2	87.5	86.2
(ii)	+ CRA	<b>92.1</b>	<b>94.6</b>	<b>87.0</b>	<b>90.6</b>	<b>86.8</b>	<b>81.6</b>	<b>94.7</b>	<b>94.1</b>
(iii)	+ SRA	87.9	92.1	83.3	87.4	84.6	78.4	89.4	87.8
(iv)	+ CA	89.6	92.8	84.2	88.2	85.2	79.7	91.2	90.1

two variations. As can be observed, the channel recurrent attention cell improves the accuracy significantly across all four datasets. This observation supports our assumption that the attention receives a performance gain from explicit modeling of the global receptive field in each slice of the feature maps.

**Effect of the Position of Channel Recurrent Attention.** The position of the channel recurrent attention block affects the information in the spatial or the channel dimensions. We want to explore the rich spatial and channel information; thus, we only consider the feature maps from the middle of the deep network as input to channel recurrent attention (*i.e.*,  $P_1$ ,  $P_2$ , and  $P_3$  in Fig. 1). The comparison is illustrated in Table 2. It shows that the system receives a better gain when adding the channel recurrent attention module at position  $P_2$ , which aligns with our motivation that more spatial information is utilized in the feature maps. The works [6, 11] also present a similar observation. When applying the attention in  $P_1$ ,  $P_2$  and  $P_3$ , the network performs at its best.

**Effect of Reduction Ratio  $1/d$  in Channel Recurrent Attention.** The ratio  $1/d$  in the embedding function  $\phi(\cdot)$  (see Fig. 2) is to reduce the channel dimensionality of the input feature maps, consequently, reducing the sequence length input to the LSTM; thus, it is an important hyper-parameter in the channel recurrent attention. Table 3 reveals that the best performance is obtained

**Table 2.** Effect of the position of channel recurrent attention across four datasets. CRA: Channel Recurrent Attention.

Position		PRID-2011		iLIDS-VID		MARS		DukeMTMC-VideoReID	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
(i)	No Attention	85.4	91.0	80.0	87.1	82.3	76.2	87.5	86.2
(ii)	+ CRA in $P_1$	89.6	92.2	85.3	88.2	85.0	80.6	92.7	92.2
(iii)	+ CRA in $P_2$	91.0	94.4	86.7	90.2	86.4	81.2	94.2	93.4
(iv)	+ CRA in $P_3$	90.3	92.6	86.0	88.4	86.1	80.8	93.5	92.7
(v)	+ CRA in $P_1 \& P_2 \& P_3$	<b>92.1</b>	<b>94.6</b>	<b>87.0</b>	<b>90.6</b>	<b>86.8</b>	<b>81.6</b>	<b>94.7</b>	<b>94.1</b>

when  $d = 16$  for small-scale datasets and  $d = 8$  for large-scale datasets. This could be due to the fact that training a network with a large amount of training samples is less prone to overfitting. Furthermore, this table also shows the fact that the LSTM has difficulties in modeling very long sequences (*e.g.* smaller  $d$  in Table 3). However, when the sequences are too short (*e.g.*,  $d = 32$ ), the channel features are compressed, such that some pattern information is lost.

**Table 3.** Effect of reduction ratio  $1/d$  in channel recurrent attention across four datasets.

Reduction Ratio		PRID-2011		iLIDS-VID		MARS		DukeMTMC-VideoReID	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
(i)	No Attention	85.4	91.0	80.0	87.1	82.3	76.2	87.5	86.2
(ii)	$d = 2$	88.7	92.1	84.0	88.7	84.8	80.2	93.4	92.8
(iii)	$d = 4$	89.8	92.6	85.6	89.1	85.2	80.3	93.9	93.4
(iv)	$d = 8$	91.0	93.2	86.3	89.4	<b>86.8</b>	<b>81.6</b>	<b>94.7</b>	94.1
(v)	$d = 16$	<b>92.1</b>	<b>94.6</b>	<b>87.0</b>	<b>90.6</b>	85.5	80.7	94.3	<b>94.3</b>
(vi)	$d = 32$	91.0	93.8	82.7	88.9	84.3	79.8	93.2	93.4

**Why using LSTM in the Channel Recurrent Attention?** In our channel recurrent attention, we use the LSTM to perform the recurrent operation for the spatial vector. We observed that once the order of the spatial vectors is fixed, the recurrent operation in the LSTM is able to learn useful information along the channel dimension. We further investigated using Bi-LSTM to replace the LSTM in the attention and evaluate its performance. Compared with LSTM, the Bi-LSTM only brings a marginal/no performance gain across different datasets, whereas it almost doubles the number of parameters and FLOPs in the attention model. Please refer to Sect. 1 of the supplementary material for details of those experiments. These empirical experimental results support the use of a regular LSTM in our attention module.

**Effect of Set Aggregation.** Table 4 shows the effectiveness of set aggregation and the effectiveness of different pooling schemes in the set aggregation block. It is clear that the individual set aggregation improves the network performance and the combination of attention modules continues to increase the performance

gain; showing that two attention modules mine complementary information in the network. Furthermore, all pooling schemes improve the results of the network, showing that the network receives gains from set aggregation. The combination of the average pooling and the max pooling scheme with non-sharing weights further shows its superiority over the individual average or max pooling schemes. This observation can be interpreted as the average and max pooled features have complementary information when encoding clip-level representations.

**Table 4.** Effect of set aggregation across four datasets. CRA: Channel Recurrent Attention, SA: Set Aggregation, †: Sharing weights, ‡: Non-sharing weights.

		PRID-2011		iLIDS-VID		MARS		DukeMTMC-VideoReID	
Method		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
(i)	No Attention	85.4	91.0	80.0	87.1	82.3	76.2	87.5	86.2
(ii)	+ CRA	92.1	94.6	87.0	90.6	86.8	81.6	94.7	94.1
(iii)	+ SA (Average & Max Pooling)	87.6	92.3	84.7	89.1	85.2	80.5	91.2	88.9
(iv)	+ CRA & SA (Avg Pooling)	94.4	95.2	87.9	91.2	87.2	82.2	95.6	95.0
(v)	+ CRA & SA (Max Pooling)	93.3	94.8	87.3	90.8	86.9	81.2	95.2	94.6
(vi)	+ CRA & SA <sup>†</sup> (Avg & Max Pooling)	95.5	96.1	88.2	92.4	87.7	82.6	95.9	95.3
(vii)	+ CRA & SA <sup>‡</sup> (Avg & Max Pooling)	<b>96.6</b>	<b>96.9</b>	<b>88.7</b>	<b>93.0</b>	<b>87.9</b>	<b>83.0</b>	<b>96.3</b>	<b>95.5</b>

### 4.3 Comparison to the State-of-the-Art Methods

To evaluate the superiority of our deep attentional network, we continue to compare our results with the current state-of-the-art approaches, shown in Table 5 and Table 6.

**PRID-2011.** On the PRID-2011 dataset, our network improves the state-of-the-art accuracy by 1.1% in R-1, compared to GLTR [51]. As for the mAP, our approach outperforms [43] by 2.4%. When compared to SCAN [46], which uses optical flow, our approach outperforms it by 1.3% in R-1.

**iLIDS-VID.** On the iLIDS-VID dataset, our approach improves the state-of-the-art mAP value by 5.2%, compared to [43]. As for the R-1 accuracy, our approach also achieves a new state-of-the-art, outperforming [33] by a comfortable 2.4%. In addition, our approach continues to outperform SCAN + optical flow [46] by 0.7% in R-1.

**MARS.** On the MARS dataset, our approach achieves state-of-the-art performances on mAP and competitive performance on the CMC curve. In particular, our approach outperforms VRSTC [50] on mAP, R-5 and R-10. It is worth mentioning that VRSTC uses a generator for data augmentation. Furthermore, when compared to other methods, we observe that our approach outperforms GLTR [51] by 1.3%/4.6% in R-1/mAP.

**Table 5.** Comparison with the SOTA methods on PRID-2011, iLIDS-VID and MARS datasets.

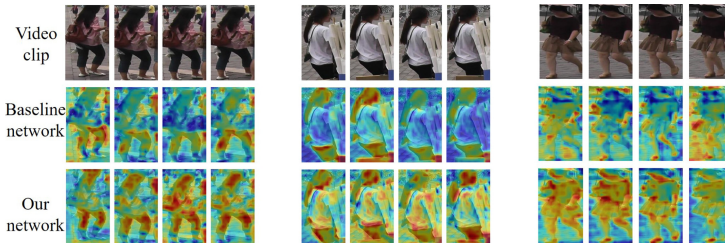
Method	Publication	PRID-2011					iLIDS-VID					MARS				
		R-1	R-5	R-10	R-20	mAP	R-1	R-5	R-10	R-20	mAP	R-1	R-5	R-10	R-20	mAP
RFA-Net [27]	ECCV'16	58.2	85.8	93.4	97.9	-	49.3	76.8	85.3	90.0	-	-	-	-	-	-
McLaughlin <i>et al.</i> [23]	CVPR'16	70.0	90.0	95.0	97.0	-	58.0	84.0	91.0	96.0	-	-	-	-	-	-
MSCAN [41]	CVPR'17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Zhou <i>et al.</i> [42]	CVPR'17	79.4	94.4	-	99.3	-	55.2	86.5	-	97.0	-	71.8	86.6	-	93.1	56.1
Chen <i>et al.</i> [43]	CVPR'18	88.6	99.1	-	-	90.9	79.8	91.8	-	-	82.6	70.6	90.0	-	97.6	50.7
+ Optical flow		93.0	99.3	100.0	100.0	94.5	85.4	96.7	98.8	99.5	87.8	81.2	92.1	-	-	69.4
QAN [44]	CVPR'17	90.3	98.2	99.3	100.0	-	68.0	86.8	-	97.4	-	86.3	94.7	-	98.2	76.1
Li <i>et al.</i> [45]	CVPR'18	93.2	-	-	-	-	80.2	-	-	-	-	73.7	84.9	-	91.6	51.7
Gao <i>et al.</i> [22]	BMVC'18	-	-	-	-	-	-	-	-	-	-	82.3	-	-	-	65.8
SCAN [46]	TIP'19	92.0	98.0	100.0	100.0	-	81.3	93.3	96.0	98.0	-	83.3	93.8	96.0	97.4	76.7
+ Optical flow		95.3	99.0	100.0	100.0	-	88.0	96.7	98.0	100.0	-	86.6	94.8	-	98.1	76.7
STIM-RRU [47]	AAAI'19	92.7	98.8	-	99.8	-	84.3	96.8	-	100.0	-	87.2	95.2	-	98.1	77.2
COSAM [8]	ICCV'19	-	-	-	-	-	79.6	95.3	-	-	-	84.4	93.2	-	96.3	72.7
STAR+Optical flow [48]	BMVC'19	93.4	98.3	100.0	100.0	-	85.9	97.1	98.9	99.7	-	84.9	95.5	-	97.9	79.9
STA [49]	AAAI'19	-	-	-	-	-	-	-	-	-	-	85.4	95.4	96.2	97.3	76.0
VRSTC [50]	CVPR'19	-	-	-	-	-	83.4	95.5	97.7	99.5	-	86.3	95.7	-	98.1	80.8
Zhao <i>et al.</i> [33]	CVPR'19	93.9	99.5	-	100.0	-	86.3	97.4	-	99.7	-	<b>88.5</b>	96.5	97.4	-	82.3
GLTR [51]	ICCV'19	95.5	<b>100.0</b>	-	-	-	86.0	98.0	-	-	-	87.0	95.4	-	98.7	78.2
Baseline	-	85.4	98.9	98.9	98.9	91.0	80.0	95.3	98.7	99.3	87.1	87.0	95.7	-	98.2	78.4
Ours	-	<b>96.6</b>	98.9	<b>100.0</b>	<b>100.0</b>	<b>96.9</b>	<b>88.7</b>	<b>97.3</b>	<b>99.3</b>	<b>100.0</b>	<b>93.0</b>	87.9	<b>96.6</b>	<b>97.5</b>	<b>98.8</b>	<b>83.0</b>

**Table 6.** Comparison with the SOTA methods on DukeMTMC-VideoReID dataset.

Method	Publication	DukeMTMC-VideoReID				
		R-1	R-5	R-10	R-20	mAP
ETAP-Net [38]	CVPR'18	83.6	94.6	-	97.6	78.3
STAR+Optical flow [48]	BMVC'19	94.0	99.0	99.3	99.7	93.4
VRSTC [50]	CVPR'19	95.0	99.1	99.4	-	93.5
STA [49]	AAAI'19	96.2	99.3	-	99.7	94.9
GLTR [51]	ICCV'19	<b>96.3</b>	99.3	-	99.7	93.7
Baseline	-	87.5	96.5	97.2	98.3	86.2
Ours	-	<b>96.3</b>	<b>99.4</b>	<b>99.7</b>	<b>99.9</b>	<b>95.5</b>

**DukeMTMC-VideoReID.** As for this new dataset, our network continues to show its superior performance (see Table 6). Our approach is superior to GLTR by 1.8% on mAP, and outperform the state-of-the-art mAP value of STA by 0.6%, and our network also achieves competitive performance on the CMC metric, outperforming the state-of-the-art on R-5, R-10 and R-20.

We visualize the feature maps from the baseline network and our channel recurrent attention network, trained on the MARS dataset. The feature maps are obtained in  $P_2$  (see Fig. 1). In Fig. 7, we observed that compared to the baseline network, our attention network highlights more areas of human bodies, which verifies the effectiveness of our network qualitatively. Please refer to the supplementary material for further visualizations.

**Fig. 7.** Visualization of feature maps. We sample three video clips from different pedestrians and visualize the feature maps.

## 5 Experiments on Single Image Pedestrian Retrieval

To show the generalisation of the proposed channel recurrent attention, we employ it in a single image pedestrian retrieval task. We select a strong baseline network from [6], and insert the channel recurrent attention after each convolutional block. The deep network is fine-tuned from ImageNet pre-training [31] and trained with the same hyper-parameter setting as in [6]. We use **CUHK01** [52] and **DukeMTMC-reID** [53] to evaluate the performance of the network. Please refer to the supplementary material for the details of the datasets. We use mAP

and the CMC curve to evaluate the performance. Table 7 and Table 8 illustrate that our approach achieves competitive results to existing state-of-the-art approaches, showing the effectiveness and generalization of our channel recurrent attention module.

**Table 7.** Comparison with the SOTA on CUHK01 dataset.

Method	Publication	CUHK01			
		R-1	R-5	R-10	R-20
Zhao <i>et al.</i> [54]	ICCV'17	75.0	93.5	95.7	97.7
Spindle Net [55]	CVPR'17	79.9	94.4	97.1	98.6
PBR [2]	ECCV'18	80.7	94.4	97.3	98.6
Baseline	-	79.3	92.7	95.8	98.2
Ours	-	<b>83.3</b>	<b>96.3</b>	<b>98.4</b>	<b>98.9</b>

**Table 8.** Comparison with the SOTA on DukeMTMC-reID dataset.

Method	Publication	DukeMTMC-reID			
		R-1	R-5	R-10	mAP
OS-Net [56]	ICCV'19	88.6	-	-	73.5
BAT-net [6]	ICCV'19	87.7	94.7	96.3	77.3
ABD-Net [57]	ICCV'19	89.0	-	-	<b>78.6</b>
Baseline	-	85.4	93.8	95.5	75.0
Ours	-	<b>89.2</b>	<b>95.6</b>	<b>96.9</b>	78.3

## 6 Conclusion

This work proposes a novel deep attentional network for task of video pedestrian retrieval. This network benefits from the developed channel recurrent attention and set aggregation modules. The channel recurrent attention module is employed for a global view to feature maps, to learn the channel and spatial pattern jointly, given a frame feature maps as input. Then the set aggregation cell continues to re-weight each frame feature and fuses them to get a compact clip representation. Thorough evaluation shows that the proposed deep network achieves state-of-the-art results across four standard video-based person re-ID datasets, and the effectiveness of each attention is further evaluated by extensive ablation studies.

## References

1. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future (2016) [arXiv:1610.02984](https://arxiv.org/abs/1610.02984) [cs.CV]
2. Suh, Y., Wang, J., Tang, S., Mei, T., Mu Lee, K.: Part-aligned bilinear representations for person re-identification. In: ECCV (2018)
3. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR (2018)
4. Zhou, J., Roy, S.K., Fang, P., Harandi, M., Petersson, L.: Cross-correlated attention networks for person re-identification. In: Image and Vision Computing (2020)
5. Wang, C., Zhang, Q., Huang, C., Liu, W., Wang, X.: Mancs: a multi-task attentional network with curriculum sampling for person re-identification. In: ECCV (2018)
6. Fang, P., Zhou, J., Roy, S.K., Petersson, L., Harandi, M.: Bilinear attention networks for person retrieval. In: ICCV (2019)
7. Chen, B., Deng, W., Hu, J.: Mixed high-order attention network for person re-identification. In: ICCV (2019)

8. Subramaniam, A., Nambiar, A., Mittal, A.: Co-segmentation inspired attention networks for video-based person re-identification. In: ICCV (2019)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
10. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: convolutional block attention module. In: ECCV (2018)
11. Wang, X., Girshick, Gupta, A., He, K.: Non-local neural networks. In: CVPR (2017)
12. Fu, J., et al.: Dual attention network for scene segmentation. In: CVPR (2019)
13. Li, W., Jafari, O.H., Rother, C.: Deep object co-segmentation. In: ACCV (2018)
14. Wang, F., et al.: Residual attention network for image classification. In: CVPR (2017)
15. Hjelm, R.D., et al.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019)
16. Gong, Shaogang., Cristani, Marco., Yan, Shuicheng, Loy, Chen Change (eds.): Person Re-Identification. ACVPR. Springer, London (2014). <https://doi.org/10.1007/978-1-4471-6296-4>
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature (2015)
18. Yi, D., Lei, Z., Li, S.Z.: Deep metric learning for person re-identification. In: ICPR (2014)
19. Bromley, J., Guyon, I., LeCun, Y., Säcker, E., Shah, R.: Signature verification using a “siamese” time delay neural network. In: NeurIPS (1993)
20. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
21. Wu, S., Chen, Y.C., Li, X., Wu, A.C., You, J.J., Zheng, W.S.: An enhanced deep feature representation for person re-identification. In: WACV (2016)
22. Gao, J., Nevatia, R.: Revisiting temporal modeling for video-based person ReID. [arXiv:1805.02104](https://arxiv.org/abs/1805.02104) (2018)
23. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: CVPR (2016)
24. Liu, H., Feng, J., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification (2016) [arXiv:1606.04404](https://arxiv.org/abs/1606.04404) [cs.CV]
25. Liu, X., et al.: Hydraplus-net: attentive deep features for pedestrian analysis. In: ICCV (2017)
26. Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., Xu, Y.: Deep-Person: learning discriminative deep features for person re-identification. Pattern Recognition (2020)
27. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X.: Person re-identification via recurrent feature aggregation. In: ECCV (2016)
28. Zhao, B., Wu, X., Feng, J., Peng, Q., Yan, S.: Diversified visual attention networks for fine-grained object classification. In: TMM (2017)
29. Paszke, A., et al.: Automatic differentiation in pytorch. In: NeurIPS (2017)
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
31. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. In: IJCV (2015)
32. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
33. Zhao, Y., Shen, X., Jin, Z., Lu, H., Hua, X.S.: Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In: CVPR (2019)
34. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. [arXiv:1708.04896](https://arxiv.org/abs/1708.04896) (2017)



35. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: *Image Analysis* (2011)
36. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. In: *TPAMI* (2016)
37. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: *ECCV* (2016)
38. Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: *CVPR* (2018)
39. Dehghan, A., Assari, S.M., Shah, M.: Gmmcp tracker: globally optimal generalized maximum multi clique problem for multiple object tracking. In: *CVPR* (2015)
40. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. In: *TPAMI* (2010)
41. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: *CVPR* (2017)
42. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In: *CVPR* (2017)
43. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: *CVPR* (2018)
44. Liu, Y., Junjie, Y., Ouyang, W.: Quality aware network for set to set recognition. In: *CVPR* (2017)
45. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: *CVPR* (2018)
46. Zhang, R., et al.: Scan: self-and-collaborative attention network for video person re-identification (2018) [arXiv:1807.05688](https://arxiv.org/abs/1807.05688) [cs.CV]
47. Liu, Y., Yuan, Z., Zhou, W., Li, H.: Spatial and temporal mutual promotion for video-based person re-identification. In: *AAAI* (2019)
48. Wu, G., Zhu, X., Gong, S.: Spatio-temporal associative representation for video person re-identification. In: *BMVC* (2019)
49. Fu, Y., Wang, X., Wei, Y., Huang, T.: STA: spatial-temporal attention for large-scale video-based person re-identification. In: *AAAI* (2019)
50. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: VRSTC: occlusion-free video person re-identification. In: *CVPR* (2019)
51. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representation for video person re-identification. In: *ICCV* (2019)
52. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: *ACCV* (2012)
53. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: *ECCVworkshop on Benchmarking Multi-Target Tracking* (2016)
54. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: *ICCV* (2017)
55. Zhao, H., et al.: Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: *CVPR* (2017)
56. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. [arXiv:1910.06827v2](https://arxiv.org/abs/1910.06827v2) (2019)
57. Chen, T., et al.: Abd-net: attentive but diverse person re-identification. In: *ICCV* (2019)