# Spatio-Temporal Associative Representation for Video Person Re-Identification

Guile Wu[1]
guile.wu@qmul.ac.uk
Xiatian Zhu[2]
eddy.zhuxt@gmail.com
Shaogang Gong[1]
s.gong@qmul.ac.uk

[1] Computer Vision Group,
School of Electronic Engineering and Computer Science,
Queen Mary University of London, London E1 4NS, UK.

[2] Vision Semantics Limited,
London E1 4NS, UK.

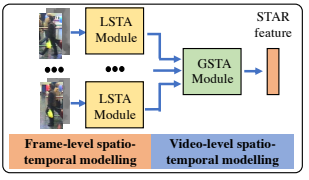
**Abstract**

Learning discriminative spatio-temporal representation is the key for solving video re-identification (re-id) challenges. Most existing methods focus on learning appearance features and/or selecting image frames, but ignore optimising the compatibility and interaction of appearance and motion attentive information. To address this limitation, we propose a novel model to learning Spatio-Temporal Associative Representation (STAR). We design local frame-level spatio-temporal association to learn discriminative attentive appearance and short-term motion features, and global video-level spatio-temporal association to form compact and discriminative holistic video representation. We further introduce a pyramid ranking regulariser for facilitating end-to-end model optimisation. Extensive experiments demonstrate the superiority of STAR against state-of-the-art methods on four video re-id benchmarks, including MARS, DukeMTMC-VideoReID, iLIDS-VID and PRID-2011.

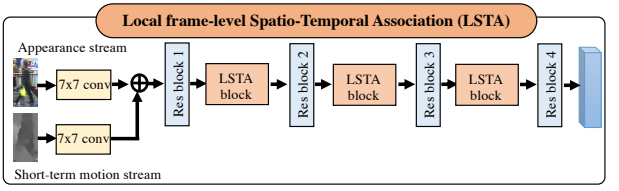
## 1 Introduction

Person re-identification (re-id), which aims to match people in images or videos across non-overlapping camera views, is a key capability for many real-world applications, such as intelligent surveillance and human computer interaction [1, 15, 36]. In general, re-id studies can be categorised as either image-based or video-based approaches [19, 42]. Most existing re-id studies are image-based methods [7, 16, 43], which focus on learning effective visual appearance features using a still image. In comparison, video re-id is closer to realistic applications, because first-hand data captured from surveillance cameras are usually videos, and more importantly, video re-id is capable of exploring richer spatial and temporal information [1, 6, 19, 45] which alleviates the misalignment and occlusion problems of image re-id in complex scenes. Therefore, learning discriminative spatio-temporal representations for video re-id is an important task for both research and applications.

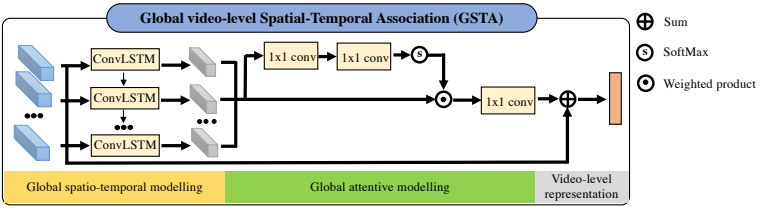
An intuitive solution for video re-id is by temporal pooling of image-level CNN appearance features [29, 32]. However, this strategy tends to be suboptimal since the quality of an individual image in each video sequence cannot be well guaranteed [15, 20], *e.g.* corrupted




(a) Overall STAR structure.




(b) Local frame-level spatio-temporal association module.




(c) Global video-level spatio-temporal association module.

Figure 1: Diagrams of spatio-temporal associative representation (STAR) learning.

images by occlusion and/or motion blur and inferior capability for modelling temporal dynamics. To make full use of spatio-temporal information, an improved strategy is to stack a Recurrent Neural Network (RNN) on top of image features as a long-term temporal feature extractor [22, 42]. However, RNN cannot incorporate spatial information during temporal learning and its long-term aggregated output is prone to being contaminated by noise especially in later steps. Inspired by attention mechanism [12, 33, 34, 37], some studies [1, 6, 15, 42, 46, 49] exploit spatio-temporal saliency to select discriminative spatial and temporal information in person videos. But such methods do not exploit mutual promotion of appearance and motion information along with attention learning therefore leading to less discriminative spatio-temporal feature representations.

In this work, we investigate the potential of jointly learning both spatio-temporal representations and attention in synergistic interaction for video re-id. We achieve this by learning a novel Spatio-Temporal Associative Representation (STAR) (Fig. 1(a)). STAR is composited of two components: (1) A *Local frame-level Spatio-Temporal Association* (LSTA) module (Fig. 1(b)) to learn discriminative per-frame appearance and short-term inter-frame motion information (optic flow). (2) A *Global video-level Spatio-Temporal Association* (GSTA) module (Fig. 1(c)) to learn compatible spatio-temporal information reinforced with long-term temporal attention. To enhance the interaction between spatial and temporal representation, we adopt Convolutional LSTM (ConvLSTM) [28] in GSTA. This however may introduces a learning difficulty when jointly optimising CNN of LSTA and ConvLSTM of GSTA. To address this issue, we further introduce a pyramid ranking regulariser to optimise the intermediate representations with deeper supervision and train the model with multiple losses in an end-to-end fashion. The contributions of this work are:

- We propose a novel end-to-end video re-id model fully exploiting appearance and motion attentive cues for learning discriminative spatio-temporal associative representations.
- We design a local frame-level spatio-temporal association module to learn attentive appearance and short-term motion information, and a global video-level spatio-temporal

association module to produce compact attentive video representations.

- We introduce a pyramid ranking regulariser for facilitating end-to-end optimisation of local and global spatio-temporal attentive representations via reinforcing intermediate features.

Extensive experiments show that the proposed STAR method outperforms state-of-the-art video re-id methods on four video re-id benchmarks, including MARS [47], DukeMTMC-VideoReID [27, 39], iLIDS-VID [36] and PRID-2011 [10].

## 2 Related Work

**Image person re-id** has been extensively investigated in the literature. Most existing image re-id methods aim to learn discriminative appearance features [7, 32, 40] and/or distance metric [9, 18, 24]. For example, Fu *et al*. [7] design a deep CNN model to learn discriminative re-id features from horizontal pyramids. Suh *et al*. [32] propose to learn part-aligned bilinear representations using a sophisticated appearance and part models. In [24], Paisitkriangkrai *et al*. introduce a learning to rank mechanism that directly optimises the evaluation measure. Although a great progress has been made in image re-id, existing image-based methods cannot achieve promising performance in videos because they only consider learning spatial appearance information without considering temporal dynamics.

**Video person re-id** attracts more attentions recently [15, 22, 23, 42, 45] due to being closer to realistic scenarios and the potential advantage of leveraging spatial and temporal information to resolve visual ambiguities including occlusion and background noise. In [17], Li *et al*. use hand-crafted local features to model motion variations and combine them with deep features for re-id. Wang *et al*. [36] propose a clip ranking approach to select discriminative video sequences for matching. Chung *et al*. [3] propose a two-stream siamese network to jointly optimise deep features and distance metric for video re-id. In our work, we learn spatio-temporal associative representations with attentive optimisation for video re-id.

**Spatio-temporal cues collaborative learning for video re-id** is one of the most effective approaches in addressing the intrinsic challenges such as occlusion and viewpoint variation. Mclaughlin *et al*. [22] propose to stack RGB frames with optical flow as inputs to a RNN model and jointly optimise the model in a siamese architecture. Liu *et al*. [19] design a refined recurrent unit for modelling temporal motion information and restoring consecutive parts from reliable historic cues to extract video-level representations. Li *et al*. [14] incorporate multi-scale 3D convolution layers into 2D CNN for spatio-temporal learning and use a two-stream network to combine spatial and temporal features. In contrast to these methods, the proposed STAR learn appearance and short-term motion information by local frame-level association and optimise video-level representations by global spatio-temporal association.

**Attentive learning for re-id** has shown its efficacy and achieved promising results in recent years [33, 37, 41]. Li *et al*. [16] propose a harmonious attention network to extract spatial attentive representations from both holistic and local regions for image re-id. Xu *et al*. [42] use spatial pyramid pooling and temporal selection to learn attentive features for video re-id. In [15], Li *et al*. present a two-stage spatio-temporal network for video re-id. They separately train a CNN model in some image re-id datasets as the deep appearance feature extractor and utilise multiple spatial and temporal models to optimise spatio-temporal gated features. In our work, we propose an end-to-end joint learning model to fully mining attentive appearance and motion cues in a synergistic interaction for more effective video re-id.

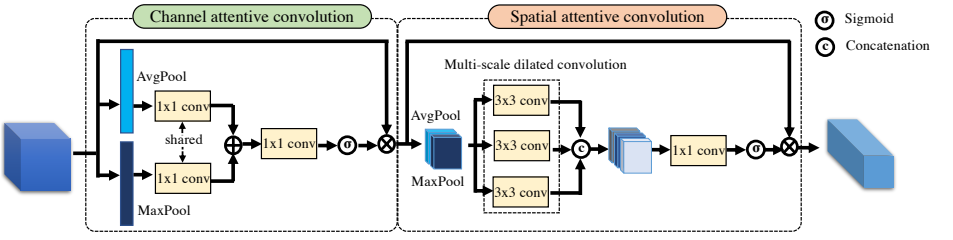



Figure 2: Structure of the local frame-level spatio-temporal association block.

# 3 Methodology

## 3.1 Framework Overview

We formulate a model to learn Spatio-Temporal Associative Representation (STAR) for video re-id. The overall structure of STAR is depicted in Fig. 1(a). We use ResNet-50 [8] as the backbone CNN. STAR contains two components: *Local frame-level Spatio-Temporal Association* (LSTA) (Fig. 1(b)) and *Global video-level Spatio-Temporal Association* (GSTA) (Fig. 1(c)). Given a video with $L$ frames, we use both RGB frames $\{I_i\}_{i=1}^{L}$ and optical flow frames $\{O_i\}_{i=1}^{L}$ as the input to LSTA to extract frame-level attentive deep features $\{f_i\}_{i=1}^{L}$:

$$\{f_i\}_{i=1}^{L} = \mathcal{F}_l(\{I_i\}_{i=1}^{L}, \{O_i\}_{i=1}^{L}) \tag{1}$$

where $\mathcal{F}_l(\cdot)$ denotes feature extraction by LSTA. This exploits both appearance and short-term motion cues (inter-frame). In contrast to two-stream action recognition [30] or action primitives dynamic programming [26], video re-id relies more on fine-grained appearance features, whilst optical flow provides motion cues for boundary regions which are variant to appearance variation [1, 22, 46]. Therefore, we separately process appearance information and short-term motion information [5] (each by a convolutional layer with kernel size $7 \times 7$), and then aggregate them as $U_i$ for the following layers:

$$U_i = \mathcal{P}(max(0, W_1 I_i)) + \mathcal{P}(max(0, W_2 O_i)) \tag{2}$$

where $\mathcal{P}(\cdot)$ is a $3 \times 3$ max pooling layer as that in ResNet-50, $W_1$ and $W_2$ are to-be-learned weights.

In addition, GSTA (Fig. 1(c)) aggregates attentively long-term spatio-temporal information at the video level and outputs the final STAR feature $\mathcal{V}$:

$$\mathcal{V} = \mathcal{F}_g(\{f_i\}_{i=1}^{L}) \tag{3}$$

where $\mathcal{F}_g(\cdot)$ denotes feature extraction in GSTA module.

Assume that there are $K = \{1, ..., k\}$ video sequences with $N = \{1, ..., n\}$ identities, we extract STAR features $\{\mathcal{V}\}_{i=1}^{k}$ for each video and use a generic distance metric (e.g. $L2$ distance) to measure their pairwise similarity for the final video re-id matching.

## 3.2 Local Frame-Level Spatio-Temporal Association Module

In LSTA module, we incorporate LSTA blocks (see Fig. 2) into the re-id model for learning local frame-level and inter-frame attentive representations, which is inspired by CBAM [38].

But unlike CBAM, we consider a block-wise design other than layer-wise for less redundancy. With a feature map $M$ from a previous residual convolutional block, we separately use average and max pooling for obtaining finer attentive feature maps $M_a^c$ and $M_m^c$. Then, we define channel attention feature map $M_c$ as:

$$M_c = M \otimes \sigma(W_4(max(0, W_3 M_a^c + b_3) + max(0, W_3 M_m^c + b_3)) + b_4) \tag{4}$$

where $\otimes$ denotes Hadamard product, $\sigma$ indicates Sigmoid function, $W_3 \in \mathbb{R}^{\frac{C}{r} \times C}$ (r is the reduction ratio), $W_4 \in \mathbb{R}^{C \times \frac{C}{r}}$, $b_3 \in \mathbb{R}^{\frac{C}{r}}$ and $b_4 \in \mathbb{R}^C$ (in this paper, unless otherwise stated, $\{W_i\}_{i=1}^9$ and $\{b_i\}_{i=3}^8$ are to-be-learned parameters). Here, the second shared convolutional layer is to facilitate the combination of two channel attentive representations. Next, we use spatial pooling to generate $M_a^s$ and $M_m^s$, and concatenate them together as $M_s$. Instead of using a large $7 \times 7$ kernel size to capture spatial context as CBAM, we leverage multi-scale dilated convolution layers [2] with $3 \times 3$ kernel size and dilated ratio $\{1, 2, 3\}$ for capturing wider-range spatial information at higher cost-effectiveness, and employ a bottleneck layer to facilitate aggregation:

$$M_g = M_c \otimes \sigma(W_6(\mathcal{F}_c(\{max(0, W_{5\_i} M_s + b_{5\_i})\}_{i=1}^3) + b_6) \tag{5}$$

where $M_g$ is the output attentive feature map and $\mathcal{F}_c(\cdot)$ denotes concatenation. We extract $\{f_i\}_{i=1}^L$ from the convolutional layer before the last pooling layer in ResNet-50.

## 3.3 Global Video-Level Spatio-Temporal Association Module

Traditional LSTM uses fully connected layers per unit, so spatial information is largely lost when aggregating video-level representations. To fully exploit global spatio-temporal cues, we adopt ConvLSTM [28] which allows to model additional associative spatio-temporal cues because of retaining convolutional structures in each unit (Fig. 1(c)):

$$\{h_i\}_{i=1}^L = \frac{1}{L_H L_W} \sum_{k=1}^{L_H} \sum_{j=1}^{L_W} \mathcal{F}_{clstm}(\{f_i\}_{i=1}^L) \tag{6}$$

where $\mathcal{F}_{clstm}$ denotes one-layer ConvLSTM, $\{h_i\}_{i=1}^L$ are hidden states, $L_H$ and $L_W$ are height and width of feature maps. Then, we use two convolutional layers with $1 \times 1$ kernel to generate 1-dimension scalar values, and use a softmax function $\phi$ to generate a temporal attentive correlation matrix $\mathcal{K}$ as:

$$\mathcal{K} = \phi(W_8 max(0, (W_7 \{h_i\}_{i=1}^L + b_7) + b_8)) \tag{7}$$

We use a bottleneck layer and extract STAR representations $\mathcal{V}$ in a residual manner to facilitate a holistic gradient optimisation:

$$\mathcal{V} = W_9 \mathcal{K} \{h_i\}_{i=1}^L + \frac{1}{L_H L_W L} \sum_{i=1}^{L} \sum_{k=1}^{L_H} \sum_{j=1}^{L_W} f_i \tag{8}$$

## 3.4 Pyramid Ranking Regulariser

Jointly training a deep attentive CNN with ConvLSTM is non-trivial, considering that video-level output from GSTA may lose some fine-grained cues from LSTA. To overcome this

problem, we reinforce the fine-grained spatial attentive cues by designing a pyramid ranking regulariser $\mathcal{R}_{pr}$. Different from [7, 19, 43], $\mathcal{R}_{pr}$ is an intermediate regulariser, directly computed using a multi-layer spatial pyramid without fully connected layers or extra model parameters. This also favourably avoids the need for more complex multi-stage training [15]. In particular, we explore a $Z$-layer spatial pyramid by dividing the feature map into $G = \{2^0, ..., 2^{Z-2}, 2^{Z-1}\}$ stripes. Formally, we compute $\mathcal{R}_{pr}$ as:

$$\mathcal{R}_{pr} = \frac{1}{BZ}\sum_{j=1}^{B}\sum_{i=1}^{Z}\mathcal{R}_{pr,G_i,j} \tag{9}$$

$$\begin{aligned}\mathcal{R}_{pr,G_i} = \frac{1}{G_i}\sum_{z=1}^{G_i} max(0, \alpha_1 + \frac{1}{L}\mathcal{D}(\sum_{i=1}^{L}(\mathcal{F}_{ta}(f_{i,z}) + \mathcal{F}_{tm}(f_{i,z})), \sum_{i=1}^{L}(\mathcal{F}_{ta}(f^p_{i,z}) + \mathcal{F}_{tm}(f^p_{i,z})))\\ - \frac{1}{L}\mathcal{D}(\sum_{i=1}^{L}(\mathcal{F}_{ta}(f_{i,z}) + \mathcal{F}_{tm}(f_{i,z})), \sum_{i=1}^{L}(\mathcal{F}_{ta}(f^n_{i,z}) + \mathcal{F}_{tm}(f^n_{i,z}))))\end{aligned} \tag{10}$$

where $\{f_{i,z}\}_{i=1}^{L}$, $\{f^p_{i,z}\}_{i=1}^{L}$ and $\{f^n_{i,z}\}_{i=1}^{L}$ are the divided $z$-th horizontal feature map of $\{f_i\}_{i=1}^{L}$ and its hard positive and negative counterparts in a mini-batch (transformed to vectors using average pooling), $B$ is the mini-batch size, $\alpha_1$ denotes a margin, $\mathcal{D}(\cdot)$ is Euclidean distance, $\mathcal{F}_{ta}(\cdot)$ and $\mathcal{F}_{tm}(\cdot)$ denotes temporal average and max pooling.

### 3.5 Optimisation Objective

To jointly optimise the proposed STAR, we consider concurrent multi-loss objective. We use softmax cross-entropy loss $\mathcal{L}_{id}$ to optimise person identity classification as:

$$\mathcal{L}_{id} = -\frac{1}{B}\sum_{i=1}^{B} y_i \log\frac{\exp(W_c\mathcal{V}_i)}{\sum_{j=1}^{N}\exp(W_n\mathcal{V}_j)} \tag{11}$$

where $y_i$ is the ground truth distribution, $W_c$ and $W_n$ are to-be-learned weights. We further employ triplet ranking loss [9] to optimise the video-level discrimination as:

$$\mathcal{L}_{trip} = \frac{1}{B}\sum_{i=1}^{B} max(0, \alpha_2 + \mathcal{D}(\mathcal{V}_i, \mathcal{V}^p_i) - \mathcal{D}(\mathcal{V}_i, \mathcal{V}^n_i)) \tag{12}$$

where $\alpha_2$ denotes a margin. The overall optimisation objective is then formulated as:

$$\mathcal{L}_{loss} = \mathcal{L}_{id} + \mathcal{L}_{trip} + \lambda\mathcal{R}_{pr} \tag{13}$$

where $\lambda$ is a weight factor. Here, $\mathcal{L}_{id}$ and $\mathcal{L}_{trip}$ are the main training objective, while $\mathcal{R}_{pr}$ is an auxiliary term to further facilitate the model optimisation (see Section 4.4 for evaluation).

## 4 Experiments

### 4.1 Datasets and Evaluation Protocol

**Datasets:** To evaluate the proposed STAR, we used four challenging video re-id benchmarks, including MARS [47], DukeMTMC-VideoReID [27, 39], PRID-2011 [10] and iLIDS-VID [36]. Example videos from the four benchmarks are shown in Fig. 3. ***MARS*** is a

(a) MARS (b) DukeMTMC (c) iLIDS-VID (d) PRID-2011

Figure 3: Example person video pairs from four video re-id benchmarks.

| Methods | Source | iLIDS-VID | | | | PRID-2011 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R1 | R5 | R10 | R20 | R1 | R5 | R10 | R20 |
| RCN [22] | CVPR16 | 58.0 | 84.0 | 91.0 | 96.0 | 70.0 | 90.0 | 95.0 | 97.0 |
| TDL [24] | CVPR16 | 56.3 | 87.6 | 95.6 | 98.3 | 56.7 | 80.0 | 87.6 | 93.6 |
| MarsCNN [17] | ECCV16 | 53.0 | 81.4 | - | 95.1 | 77.3 | 93.5 | - | 99.3 |
| TSSCN [3] | ICCV17 | 60.0 | 86.0 | 93.0 | 97.0 | 78.0 | 94.0 | 97.0 | 99.0 |
| ASTPN [27] | ICCV17 | 62.0 | 86.0 | 94.0 | 98.0 | 77.0 | 95.0 | 99.0 | 99.0 |
| STRN [29] | CVPR17 | 55.2 | 86.5 | - | 97.0 | 79.4 | 94.4 | - | 99.3 |
| QAN [20] | CVPR17 | 68.0 | 86.8 | 95.4 | 97.4 | 90.3 | 98.2 | 99.3 | **100** |
| RQEN [31] | AAAI18 | 76.1 | 92.9 | 97.5 | 99.3 | 92.4 | **98.8** | 99.6 | **100** |
| EIBC [28] | CVPR18 | 44.7 | 57.3 | 63.3 | 68.7 | 70.9 | 78.7 | 82.7 | 87.3 |
| SDM [45] | CVPR18 | 60.2 | 84.7 | 91.7 | 97.4 | 85.2 | 97.1 | 98.9 | 99.6 |
| STAN [15] | CVPR18 | 80.2 | - | - | - | **93.2** | - | - | - |
| Snippet [1] | CVPR18 | **85.4** | 96.7 | **98.8** | **99.5** | 93.0 | **99.3** | **100** | **100** |
| STMP [19] | AAAI19 | 84.3 | **96.8** | - | **99.5** | 92.7 | **98.8** | - | 99.8 |
| ResNet-50 | Backbone | 69.0 | 89.1 | 94.1 | 97.3 | 86.3 | 97.8 | 99.3 | 99.8 |
| STAR | Ours | **85.9** | **97.1** | **98.9** | **99.7** | **93.4** | 98.3 | **100** | **100** |

Table 1: Comparisons with state-of-the-art video re-id methods on iLIDS-VID and PRID-2011. The best results are shown in **red bold**, while second-best in **blue bold**.

large-scale video re-id benchmark with 1,261 person identities and 20,478 tracklets captured from 6 outdoor camera views. We follow the original evaluation splits [47], *i.e.* using 625 identities with 8,298 tracklets for training, and the remaining 636 identities with 12,180 tracklets for testing. ***DukeMTMC-VideoReID*** is a recently released large-scale video re-id benchmark. There are 1,812 person identities with 4,832 tracklets in this benchmark. Following [39], we selected 702/702 identities for training/testing, with 402 identities as distractors. There are 2,196 tracklets for training and 2,636 tracklets for testing and distractors. ***iLIDS-VID*** consists of 300 person identities with 600 tracklets captured by two camera views. We used all identities and tested 10 standard random splits of 50% training and 50% testing [15, 36, 49]. ***PRID-2011*** contains 934 identities with 1,134 tracklets captured by two camera views, but only the first 200 identities appear in both views. We followed the previous studies [19, 20, 36, 49] by randomly splitting the dataset into 10 splits for training and testing.

**Evaluation Metrics:** For facilitating comparison, we used Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as the performance evaluation metrics.

## 4.2 Implementation Details

We employed ResNet-50 [8] as the backbone CNN model, which was pretrained on ImageNet [4]. We resized both RGB and optical flow frames (computed using TV-L1 [25]) to

| Methods | Source | MARS | | | | | DukeMTMC-VideoReID | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | R1 | R5 | R10 | R20 | mAP | R1 | R5 | R10 | R20 |
| MarsCNN [17] | ECCV16 | 49.3 | 68.3 | 82.6 | - | 89.4 | - | - | - | - | - |
| GOG† [21] | CVPR16 | 24.9 | 42.0 | - | - | - | 52.4 | 58.8 | - | - | - |
| ASTPN [42] | ICCV17 | - | 44.0 | 70.0 | 74.0 | 81.0 | - | - | - | - | - |
| STRN [49] | CVPR17 | 50.7 | 70.6 | 90.0 | - | **97.6** | - | - | - | - | - |
| RQEN [31] | AAAI18 | 51.7 | 73.7 | 84.9 | - | 91.6 | - | - | - | - | - |
| DAL⋆ [35] | arXiv18 | 65.0 | 75.4 | - | - | - | **83.5** | **87.0** | - | - | - |
| SDM [45] | CVPR18 | - | 71.2 | 85.7 | 91.8 | 94.3 | - | - | - | - | - |
| EUG⋆ [39] | CVPR18 | 67.4 | 80.8 | 92.1 | - | 96.1 | 78.3 | 83.6 | 94.6 | - | 97.6 |
| DuATM [29] | CVPR18 | 67.7 | 81.2 | 92.5 | - | - | - | - | - | - | - |
| STAN [15] | CVPR18 | 65.8 | 82.3 | - | - | - | - | - | - | - | - |
| Snippet [1] | CVPR18 | **76.1** | **86.3** | **94.7** | - | **98.2** | - | - | - | - | - |
| PABR [32] | ECCV18 | 72.2 | 83.0 | 92.8 | **95.0** | 96.8 | - | - | - | - | - |
| STMP [19] | AAAI19 | 72.7 | 84.4 | 93.2 | - | 96.3 | - | - | - | - | - |
| ResNet-50 | Backbone | 62.5 | 76.7 | 90.0 | 92.7 | 95.8 | 79.7 | 82.2 | **95.2** | **97.2** | **98.6** |
| STAR | Ours | **76.0** | **85.4** | **95.4** | **96.2** | 97.3 | **93.4** | **94.0** | **99.0** | **99.3** | **99.7** |

Table 2: Comparisons with state-of-the-art video re-id methods on MARS and DukeMTMC-VideoReID. ⋆Supervised EUG and DAL. †Results reported in [35].

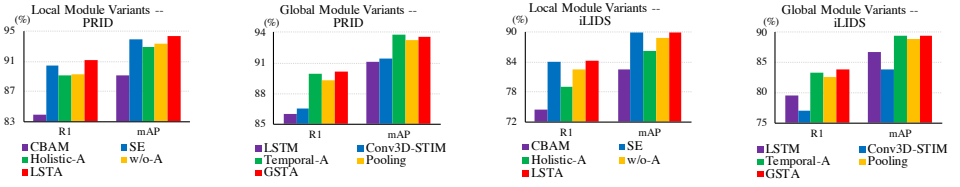
| Component | iLIDS | | PRID | | MARS | | Duke | |
|---|---|---|---|---|---|---|---|---|
| | R1 | R5 | R1 | R5 | R1 | R5 | R1 | R5 |
| Baseline-{ResNet50-ID} | 69.0 | 89.1 | 86.3 | 97.8 | 76.7 | 90.0 | 82.2 | 95.2 |
| Baseline-{CBAM-Multi-loss} | 74.7 | 92.8 | 83.9 | 95.6 | 83.7 | 93.9 | 91.7 | 98.6 |
| LSTA | 84.3 | 96.6 | 91.2 | 98.2 | 84.7 | 93.9 | 93.4 | 98.9 |
| LSTA + GSTA | 85.1 | 96.4 | 92.2 | 99.1 | 84.9 | 95.0 | 93.7 | 99.0 |
| LSTA + GSTA + $\mathcal{R}_{pr}$ | 85.9 | 97.1 | 93.4 | 98.3 | 85.4 | 95.4 | 94.0 | 99.0 |

Table 3: Evaluating component effectiveness.

$256 \times 128$. Random horizontal flip and translation were used for training data augmentation. We used Adam optimiser [13] with initial learning rate $5e$-4 and additional coefficients $\{\beta_1 = 0.9, \beta_2 = 0.999\}$. The learning rate decays by 10 times after 150 training epochs. We empirically set $r = 16$ in Eq. (4) and set $\lambda = 0.1$ in Eq. (13). In Eq. (10) and Eq. (12), $\alpha_1$ and $\alpha_2$ were both set to 0.4. We set spatial pyramid layers $Z = 3$, so $G = \{1,2,4\}$. The dimension of STAR feature was set to 2048. We set $B = 16$ and $L = 10$ (random sampling) for training, and in testing, all frames in each video were used to compute STAR features for matching.

## 4.3 Comparisons with the State-of-the-Art

Table 1 and Table 2 compare the performance of the proposed STAR with state-of-the-art methods on the four benchmarks. Here, backbone model is ResNet-50 which uses RGB and flow streams as the input and use identity loss as training objective. Overall, STAR achieves the best performance suggesting the efficacy of the proposed spatial-temporal feature and attentive joint learning method. On ***iLIDS-VID*** (see Table 1), STAR performs best consistently and outperforms the state-of-the-art by 0.5%, 0.3%, 0.1% and 0.2% on rank-1, rank-5, rank-10 and rank-20 accuracy, respectively. On ***PRID-2011*** (see Table 1), STAR




(a) (b) (c) (d)

Figure 4: Evaluating component variants.

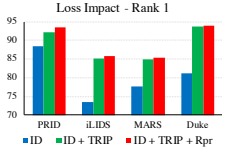



Figure 5: Loss impact.

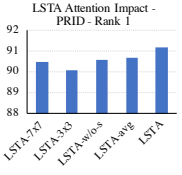



Figure 6: Attention impact.

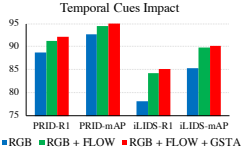



Figure 7: Temporal impact.

ranks the first on rank-1 (93.4%), rank-10 (100%) and rank-20 accuracy (100%). On ***MARS*** (see Table 2), STAR achieves second-best performance in terms of mAP (76.0%) and rank-1 accuracy (85.4%), and improves state-of-the-arts by 0.7% and 1.2% on rank-5 and rank-10 accuracy, respectively. On ***DukeMTMC-VideoReID*** (see Table 2), STAR significantly outperforms the state-of-the-art methods (achieves 93.4% and 94.0% in terms of mAP and rank-1 accuracy, respectively).

## 4.4 Ablation Studies

To further validate the proposed STAR, we conduct detailed ablation analysis as below.

**Component Effectiveness Evaluation.** In Table 3, the first two rows are baseline models: ResNet-50 with identity loss and ResNet-50 with CBAM [38] and multi-loss. Overall, LSTA, LSTA+GSTA and LSTA+GSTA+$\mathcal{R}_{pr}$ perform better than both baselines. As shown in the last three rows, GSTA can further improve the performance beyond LSTA, while LSTA with GSTA and $\mathcal{R}_{pr}$ (*i.e.* the full STAR model) achieves the best performance. These verify the positive influence of all three STAR components.

**Component Variants Comparison.** To further verify the proposed method, we investigate additional component design variants. For fair and focused comparison, we use LSTA *w/o* attention as the backbone. As shown in Fig. 4(a) and 4(c), we compare the proposed LSTA with CBAM [38], SE [12], holistic attention (two linear transforms and SoftMax as [15]), and no attention. The results show that LSTA performs better than its counterparts. As shown in Fig. 4(b) and 4(d), we employ various global aggregation modules, including GSTA, LSTM [10], Conv3D-STIM [19], holistic temporal attention (one linear transform and SoftMax as [15]), and pooling. Overall, GSTA achieves better performance compared with other variants in the proposed architecture.

**Loss Impact Evaluation.** As shown in Fig. 5, STAR trained with single $\mathcal{L}_{id}$ performs worst, while STAR with $\mathcal{L}_{id}+\mathcal{L}_{trip}$ performs significantly better. Besides, $\mathcal{R}_{pr}$ can further optimise the STAR model to achieve the best performance.

**LSTA Attention Impact Evaluation.** As shown in Fig. 6, to evaluate the improvement of CBAM in LSTA, we compare LSTA with CBAM, LSTA with $7 \times 7$ kernel as CBAM, LSTA with $3 \times 3$ kernel instead of dilated convolution, and LSTA *w/o* spatial attentive convolution. Overall, the proposed LSTA performs the best.

**Temporal Cues Impact Evaluation.** In Fig. 7, RGB, FLOW and GSTA denote appearance cues, short-term temporal cues and long-term temporal cues, respectively. For better evaluation, $\mathcal{R}_{pr}$ is not used here. It can be seen that short-term cues and long-term temporal cues are beneficial to extract finer features for video re-id and bring better performance.

# 5 Conclusions

In this work, we propose to learn spatio-temporal associative representations along with attention in synergistic compatibility for video person re-identification. Specifically, we design a novel end-to-end architecture to simultaneously learn appearance and short-term motion attentive cues by local spatio-temporal association and learn the long-term coherent dynamics of final video representations by global video-level spatio-temporal association. We further introduce a pyramid ranking regulariser for facilitating local and global spatio-temporal joint learning. Extensive experiments on four video re-id benchmarks show the superiority of the proposed model against state-of-the-art methods. We further conduct detailed model component analysis for verifying our model formulation considerations.

# Acknowldgements

This work is supported by Queen Mary University of London Principal's Scholarship, Vision Semantics Limited, Alan Turing Institute Turing Fellowship, and Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149).

# References

[1] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, 2018.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[3] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, 2017.

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

[5] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017.

[6] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. STA: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, 2019.

[7] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

[10] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *NEURAL COMPUT.*, 1997.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[14] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, 2019.

[15] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

[16] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[17] Youjiao Li, Li Zhuo, Jiafeng Li, Jing Zhang, Xi Liang, and Qi Tian. Video-based person re-identification by deep feature guided pooling. In *CVPRW*, 2017.

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

[19] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019.

[20] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017.

[21] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016.

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

[23] Deqiang Ouyang, Jie Shao, Yonghui Zhang, Yang Yang, and Heng Tao Shen. Video-based person re-identification via self-paced learning and deep reinforcement learning framework. In *ACM MM*, 2018.

[24] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015.

[25] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. TV-L1 optical flow estimation. *IPOL*, 2013.

[26] Víctor Ponce-López, Hugo Jair Escalante, Sergio Escalera, and Xavier Baró. Gesture and action recognition by evolved dynamic subgestures. In *BMVC*, 2015.

[27] 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*, 2016.

[28] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015.

[29] 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*, 2018.

[30] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[31] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018.

[32] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[34] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017.

[35] Menglin Wang, Baisheng Lai, Zhongming Jin, Xiaojin Gong, Jianqiang Huang, and Xiansheng Hua. Deep active learning for video-based person re-identification. *arXiv preprint arXiv:1812.05785*, 2018.

[36] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[38] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018.

[39] 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*, 2018.

[40] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018.

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[42] 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*, 2017.

[43] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE TIP*, 2019.

[44] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *CVPR*, 2016.

[45] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *CVPR*, 2018.

[46] Ruimao Zhang, Hongbin Sun, Jingyu Li, Yuying Ge, Liang Lin, Ping Luo, and Xiaogang Wang. SCAN: Self-and-collaborative attention network for video person re-identification. *IEEE TIP*, 2019.

[47] 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*, 2016.

[48] Jiahuan Zhou, Bing Su, and Ying Wu. Easy identification from better constraints: Multi-shot person re-identification from reference constraints. In *CVPR*, 2018.

[49] 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*, 2017.