Spatio-Temporal Associative Representation
for Video Person Re-Identification

Guile Wu

Xiatian Zhu

Shaogang Gong

1 Computer Vision Group,
School of Electronic Engineering and
Computer Science,
Queen Mary University of London, Lon-
don 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 in-
teraction 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 atten-
tive appearance and short-term motion features, and global video-level spatio-temporal
association to form compact and discriminative holistic video representation. We fur-
ther introduce a pyramid ranking regulariser for facilitating end-to-end model optimisa-
tion. 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 ap-
plications, 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 infor-
mation [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 appear-
ance 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 [34, 35], e.g. corrupted

© 2019. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
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 composed 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 introduce 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

---

**Figure 1:** Diagrams of spatio-temporal associative representation (STAR) learning.
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 [22], 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 [3], 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.
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 \(\{f_i\}_{i=1}^{L}\):

\[
\{f_i\}_{i=1}^{L} = \mathcal{F}_i(\{I_i\}_{i=1}^{L}, \{O_i\}_{i=1}^{L})
\]  

(1)

where \(\mathcal{F}_i(\cdot)\) 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:

\[
U_i = \mathcal{P}(\max(0, W_1 I_i)) + \mathcal{P}(\max(0, W_2 O_i))
\]

(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 \(V\):

\[
V = \mathcal{F}_g(\{f_i\}_{i=1}^{L})
\]

(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 \(\{V\}_{i=1}^{k}\) for each video and use a generic distance metric (e.g. \(L_2\) 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].

Figure 2: Structure of the local frame-level spatio-temporal association block.
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 $M^c_a$ and $M^c_m$. Then, we define channel attention feature map $M^c$ as:

$$M^c = M \otimes \sigma (W_4 (\max(0, W_3 M^c_a + b_3) + \max(0, W_3 M^c_m + b_3)) + b_4)$$

(4)

where $\otimes$ denotes Hadamard product, $\sigma$ indicates Sigmoid function, $W_3 \in \mathbb{R}^{C \times C}$ ($r$ is the reduction ratio), $W_4 \in \mathbb{R}^{C \times C}$, $b_3 \in \mathbb{R}^C$ and $b_4 \in \mathbb{R}^C$ (in this paper, unless otherwise stated, $\{W_i\}_i^9 = 1$ and $\{b_i\}_i^3$ 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 $M^s_a$ and $M^s_m$, 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 \cite{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:

$$M_g = M^c \otimes \sigma (F_c (\{\max(0, W_5 h^s_i) + b_5\}_i^L) + b_6)$$

(5)

where $M_g$ is the output attentive feature map and $F_c(\cdot)$ denotes concatenation. We extract $\{f_i\}_i^L$ 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 \cite{28} which allows to model additional associative spatio-temporal cues because of retaining convolutional structures in each unit (Fig. 1(c)):

$$\{h_i\}_i^L = \frac{1}{L_H L_W} \sum_{k=1}^{L_H} \sum_{j=1}^{L_W} F_{clstm}(\{f_i\}_i^L)$$

(6)

where $F_{clstm}$ denotes one-layer ConvLSTM, $\{h_i\}_i^L$ 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 $K$ as:

$$K = \phi (W_8 \max(0, (W_7 \{h_i\}_i^L + b_7) + b_8))$$

(7)

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

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

(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 $R_{pr}$. Different from [1, 13, 43], $R_{pr}$ 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 $R_{pr}$ as:

$$R_{pr} = \frac{1}{BZ} \sum_{j=1}^{B} \sum_{i=1}^{Z} R_{pr, G_i, j}$$

(9)

$$R_{pr, G_i} = \frac{1}{G_i} \sum_{z=1}^{G_i} \max(0, \alpha_1 + \frac{1}{L} \mathcal{D}(\sum_{i=1}^{L} (\mathcal{F}_{ia}(f_{i,z}) + \mathcal{F}_{im}(f_{i,z})), \sum_{i=1}^{L} (\mathcal{F}_{ia}(f'_{i,z}) + \mathcal{F}_{im}(f'_{i,z}))))$$

$$- \frac{1}{L} \mathcal{D}(\sum_{i=1}^{L} (\mathcal{F}_{ia}(f_{i,z}) + \mathcal{F}_{im}(f_{i,z})), \sum_{i=1}^{L} (\mathcal{F}_{ia}(f''_{i,z}) + \mathcal{F}_{im}(f''_{i,z}))))$$

(10)

where $\{f_{i,z}\}_{i=1}^{L}$, $\{f'_{i,z}\}_{i=1}^{L}$ and $\{f''_{i,z}\}_{i=1}^{L}$ 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, $\mathcal{F}_{ia}(\cdot)$ and $\mathcal{F}_{im}(\cdot)$ 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 $L_{id}$ to optimise person identity classification as:

$$L_{id} = -\frac{1}{B} \sum_{i=1}^{B} y_i \log \frac{\exp(W_c v_i)}{\sum_{j=1}^{N} \exp(W_n v_j)}$$

(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 [13] to optimise the video-level discrimination as:

$$L_{trip} = \frac{1}{B} \sum_{i=1}^{B} \max(0, \alpha_2 + \mathcal{D}(v_i, v'_{i}) - \mathcal{D}(v_i, v''_{i}))$$

(12)

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

$$L_{loss} = L_{id} + L_{trip} + \lambda R_{pr}$$

(13)

where $\lambda$ is a weight factor. Here, $L_{id}$ and $L_{trip}$ are the main training objective, while $R_{pr}$ 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 [13, 27], PRID-2011 [10] and iLIDS-VID [36]. Example videos from the four benchmarks are shown in Fig. 3. **MARS** is a
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 \cite{7}, \textit{i.e.} using 625 identities with 8,298 tracklets for training, and the remaining 636 identities with 12,180 tracklets for testing. \textit{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 \cite{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. \textit{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 \cite{16,20,36,49}. \textit{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 \cite{16,19,36,49} by randomly splitting the dataset into 10 splits for training and testing.

\textbf{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 \cite{8} as the backbone CNN model, which was pretrained on ImageNet \cite{22}. We resized both RGB and optical flow frames (computed using TV-L1 \cite{25}) to
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].

<table>
<thead>
<tr>
<th>Component</th>
<th>iLIDS R1</th>
<th>iLIDS R5</th>
<th>PRID R1</th>
<th>PRID R5</th>
<th>MARS R1</th>
<th>MARS R5</th>
<th>Duke R1</th>
<th>Duke R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-{ResNet50-ID}</td>
<td>69.0</td>
<td>89.1</td>
<td>86.3</td>
<td>97.8</td>
<td>76.7</td>
<td>90.0</td>
<td>82.2</td>
<td>95.2</td>
</tr>
<tr>
<td>Baseline-{CBAM-Multi-loss}</td>
<td>74.7</td>
<td>92.8</td>
<td>83.9</td>
<td>95.6</td>
<td>83.7</td>
<td>93.9</td>
<td>91.7</td>
<td>98.6</td>
</tr>
<tr>
<td>LSTA</td>
<td>84.3</td>
<td>96.6</td>
<td>91.2</td>
<td>98.2</td>
<td>84.7</td>
<td>93.9</td>
<td>93.4</td>
<td>98.9</td>
</tr>
<tr>
<td>LSTA + GSTA</td>
<td>85.1</td>
<td>96.4</td>
<td>92.2</td>
<td>99.1</td>
<td>84.9</td>
<td>95.0</td>
<td>93.7</td>
<td>99.0</td>
</tr>
<tr>
<td>LSTA + GSTA + R_{pr}</td>
<td>85.9</td>
<td>97.1</td>
<td>93.4</td>
<td>98.3</td>
<td>85.4</td>
<td>95.4</td>
<td>94.0</td>
<td>99.0</td>
</tr>
</tbody>
</table>

Table 3: Evaluating component effectiveness.

256 × 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 \textit{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 \textit{PRID-2011} (see Table 1), STAR...
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+\( R_{pr} \) 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 \( R_{pr} \) (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 [11], 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, \( R_{pr} \) 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, $R_{pr}$ 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.

Acknowledgements

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


