Supplementary Material for paper "Co-segmentation Inspired Attention Networks for Video-based Person Re-identification"

Arulkumar Subramaniam, Athira Nambiar, Anurag Mittal
Department of Computer Science and Engineering,
Indian Institute of Technology Madras
{aruls, anambiar, amittal}@cse.iitm.ac.in

Pseudo-code of COSAM module

The COSAM layer makes use of two subroutines *viz.* COSAM-SPATIAL-ATTENTION() & COSAM-CHANNEL-ATTENTION() to achieve the notion of co-segmentation as shown in the Algorithm 1.

Algorithm 1 COSAM Layer

Input: Frame-level feature maps after L^{th} CNN Block, $\mathbf{F^L} = \{F_i^L\}_{i=1}^N$
where $sizeof(\mathbf{F^L}) = N \times D_L \times H_L \times W_L$,
i = Index of the frame in the video, N=Total number of frames,
D_L=Number of channels, H_L=Height, W_L=Width

Output: Co-segmented feature maps $\mathbf{F^L_{cosam}} = \textbf{COSAM}\left(\{F_i^L\}_{i=1}^N\right)$ with same dimension as of $\mathbf{F^L}$

---

procedure COSAM-SPATIAL-ATTENTION($\mathbf{F}$) ▷ $\mathbf{F}$=Input feature maps to the subroutine

$\mathbf{F_R} = \texttt{conv\_1x1}(\mathbf{F})$ ▷ Dimension reduction step : $D_L \rightarrow D_R$

$\mathbf{F_R} = \mathbf{F_R}.view\_as(N \times D_R \times (H_L * W_L))$ ▷ Flatten the height and width dimension

Calculate cost volume for n^{th} frame feature map

$$\text{Cost Volume } \mathbf{C_n}(i,j) = \{NCC\left(\mathbf{F^{(i,j)}_{R_n}}, \mathbf{F^{(h,w)}_{R_m}}\right) \mid (1 \le m \le N, m \neq n) \; (1 \le h \le H_L) \; (1 \le w \le W_L)\}$$

$$\triangleright\; sizeof(\mathbf{C}) = N \times ((N-1) * H_L * W_L) \times H_L \times W_L$$

Spatial mask $\mathbf{M_{spatial}} = \texttt{Sigmoid(conv\_1x1}(\mathbf{C}))$ ▷ Summarization step: $((N-1) * H_L * W_L) \rightarrow 1$

▷ $sizeof(\mathbf{M_{spatial}}) = N \times 1 \times H_L \times W_L$

Spatial attended features $\mathbf{F_{spatial}} = \mathbf{M_{spatial}} \bigodot \mathbf{F}$ ▷ $\bigodot$ = point-wise multiplication

return $\mathbf{F_{spatial}}$

---

procedure COSAM-CHANNEL-ATTENTION($\mathbf{F}$) ▷ $\mathbf{F}$=Input feature maps to the subroutine

$\mathbf{F_{GAP}} = GAP(\mathbf{F})$ ▷ Global average pooling (GAP), $sizeof(\mathbf{F_{GAP}}) = N \times D_L$

Individual channel mask $\mathbf{M_{ind}} = Sigmoid(MLP(\mathbf{F_{GAP}}))$ ▷ $sizeof(\mathbf{M_{ind}}) = N \times D_L$

Common channel activation mask $\mathbf{M_{channel}} = \texttt{avg-pool}(\mathbf{M_{ind}})$ ▷ $sizeof(\mathbf{M_{channel}}) = 1 \times D_L$

Masked channel features $\mathbf{F_{channel}} = \mathbf{M_{channel}} \bigotimes \mathbf{F}$ ▷ $\bigotimes$ = Channel-wise multiplication

return $\mathbf{F_{channel}}$

---

procedure COSAM($\mathbf{F^L}$)

$\mathbf{F^L_{spatial}}$ = COSAM-SPATIAL-ATTENTION($\mathbf{F^L}$) ▷ invoke spatial attention procedure

$\mathbf{F^L_{cosam}}$ = $\mathbf{F^L_{spatial}}$ + COSAM-CHANNEL-ATTENTION($\mathbf{F^L_{spatial}}$) ▷ invoke channel attention procedure

return $\mathbf{F^L_{cosam}}$

1. Ablation study (continued)

In this section, the best performing model *SE-ResNet50+COSAM*$_{4,5}$+ *TP*$_{avg}$ is further subjected to ablation studies with respect to (a) Selection of frames, (b) Cross-dataset performances and (c) Effect of spatial vs. Channel attention in COSAM.

1.1. Selection of frames

We perform experiments with two schemes of frame selection during training namely: 1) *Sequential*: a continuous sequence of N frames are selected, 2) *Random*: random sampled N frames from the whole video sequence. The quantitative results are shown in Table 1. It is observed that the performance of training using randomly sampled frames is inferior to the performance of training using sequentially sampled frames.

frame selection	MARS				DukeMTMC-VideoReID			
	mAP	R1	R5	R20	mAP	R1	R5	R20
sequential	**79.9**	**84.9**	**95.5**	**97.9**	**94.1**	**95.4**	**99.3**	**99.8**
random	77.5	83.3	93.6	97.0	90.2	90.6	98.3	99.6

Table 1. Evaluation of the influence of frame selection on Re-ID performance of the best performing model *SE-ResNet50+COSAM*$_{4,5}$+*TP*$_{avg}$.

1.2. Cross-dataset test performance

We analyze the cross-dataset performance of the best performing model (*SE-ResNet50+COSAM*$_{4,5}$+ *TP*$_{avg}$), against a base-model without using COSAM layer. The results are shown in Table 2. By training on MARS dataset and testing on DukeMTMC-VideoReID dataset, the former model outperforms the latter by 2.8% mAP and 3.5% CMC Rank-1. Similar performance of COSAM-based model is observed while training on DukeMTMC-VideoReID dataset and testing on MARS dataset as well (Improvement of 0.9% in mAP and 0.7% in CMC Rank-1).

	Train set	Test set	mAP	R1	R5	R20
No COSAM	MARS	DukeMTMC	32.0	33.3	53.3	67.1
COSAM$_{4,5}$	MARS	DukeMTMC	**34.8**	**36.8**	**54.1**	**67.9**
No COSAM	DukeMTMC	MARS	25.0	41.7	54.4	65.3
COSAM$_{4,5}$	DukeMTMC	MARS	**25.9**	**42.4**	**56.0**	**65.8**

Table 2. Cross-dataset performance of the best performing model with *SE-ResNet50* as the feature extractor and *TP*$_{avg}$ as the temporal aggregation layer. Here *DukeMTMC* = DukeMTMC-VideoReID.

Attention layer	MARS				DukeMTMC-VideoReID			
	mAP	R1	R5	R20	mAP	R1	R5	R20
Only spatial attention	78.8	84.1	94.9	97.7	93.6	93.9	99.0	**99.9**
Only Channel attention	79.0	84.3	95.0	97.8	93.8	94.4	99.1	99..7
Both spatial and channel	**79.9**	**84.9**	**95.5**	**97.9**	**94.1**	**95.4**	**99.3**	99.8

Table 3. Evaluation of the influence of Co-segmentation based attention layers on Re-ID performance of the best performing model *SE-ResNet50+COSAM*$_{4,5}$+ *TP*$_{avg}$.

1.3. Spatial *vs.* Channel attention

To evaluate the effect of the attention steps in the COSAM layer, we investigate the individual attention steps (Spatial and Channel). The results in Table 3 shows that the inclusion of both spatial and channel attention steps together achieves superior performance.

2. Location of COSAM module (multiple COSAM)

In addition to the Table 3 in the main paper, we experiment by inserting COSAM layer simultaneously after multiple CNN blocks and report the performance in Table 4. It can be observed that by adding our plug-and-play COSAM layer at multiple positions in the network (especially at deeper layers), the system performance keeps on improving. Keeping the computation overload and memory efficiency in mind, COSAM$_{4,5}$ variant of the model with feature extractor as SE-ResNet50 is selected to be the *de facto* model for other experiments.

	COSAM$_i$	MARS				DukeMTMC-VideoReID			
		mAP	R1	R5	R20	mAP	R1	R5	R20
ResNet50	No COSAM [1]	75.8	83.1	92.8	96.8	92.9	93.6	99.0	99.7
	COSAM$_2$	68.3	77.7	90.1	96.1	88.9	90.2	98.4	99.0
	COSAM$_3$	76.9	82.7	**94.3**	97.3	93.6	94.0	98.7	**99.9**
	COSAM$_4$	76.8	82.9	94.2	97.1	93.8	**94.7**	98.7	99.7
	COSAM$_5$	76.6	82.8	93.9	97.2	93.2	93.7	98.4	**99.9**
	COSAM$_{3,4}$	76.4	83.4	93.9	97.1	93.7	94.4	99.1	99.4
	COSAM$_{3,5}$	76.9	**83.7**	94.0	97.3	93.0	93.7	99.0	99.7
	COSAM$_{4,5}$	**77.2**	**83.7**	94.1	**97.5**	**94.0**	94.4	**99.1**	**99.9**
	COSAM$_{3,4,5}$	76.6	83.2	93.7	97.3	93.1	93.6	98.7	99.4
SE-ResNet50	No COSAM	78.3	84.0	95.2	97.1	93.5	93.7	99.0	99.7
	COSAM$_2$	67.0	77.9	90.4	94.9	92.2	94.0	98.9	99.7
	COSAM$_3$	79.5	85.0	94.7	97.8	93.6	94.7	99.0	**99.9**
	COSAM$_4$	79.8	84.9	95.4	97.8	94.0	**95.4**	99.0	**99.9**
	COSAM$_5$	**79.9**	84.5	**95.7**	97.9	93.9	94.9	99.1	**99.9**
	COSAM$_{3,4}$	79.5	84.8	94.7	97.6	93.7	94.7	98.7	99.7
	COSAM$_{3,5}$	79.8	**85.2**	95.5	98.0	93.9	94.2	99.3	99.9
	COSAM$_{4,5}$	**79.9**	84.9	95.5	**97.9**	**94.1**	**95.4**	**99.3**	99.8
	COSAM$_{3,4,5}$	**80.5**	**85.2**	95.5	**98.0**	**94.1**	**95.4**	**99.3**	**99.9**

Table 4. Evaluation of the backbone feature extractors with COSAM and temporal aggregation layer as TP$_{avg}$. Here, *COSAM*$_i$ = plugging in COSAM layer after i^{th} *CNN* block.

3. State-of-the-art comparison on iLIDS-VID dataset

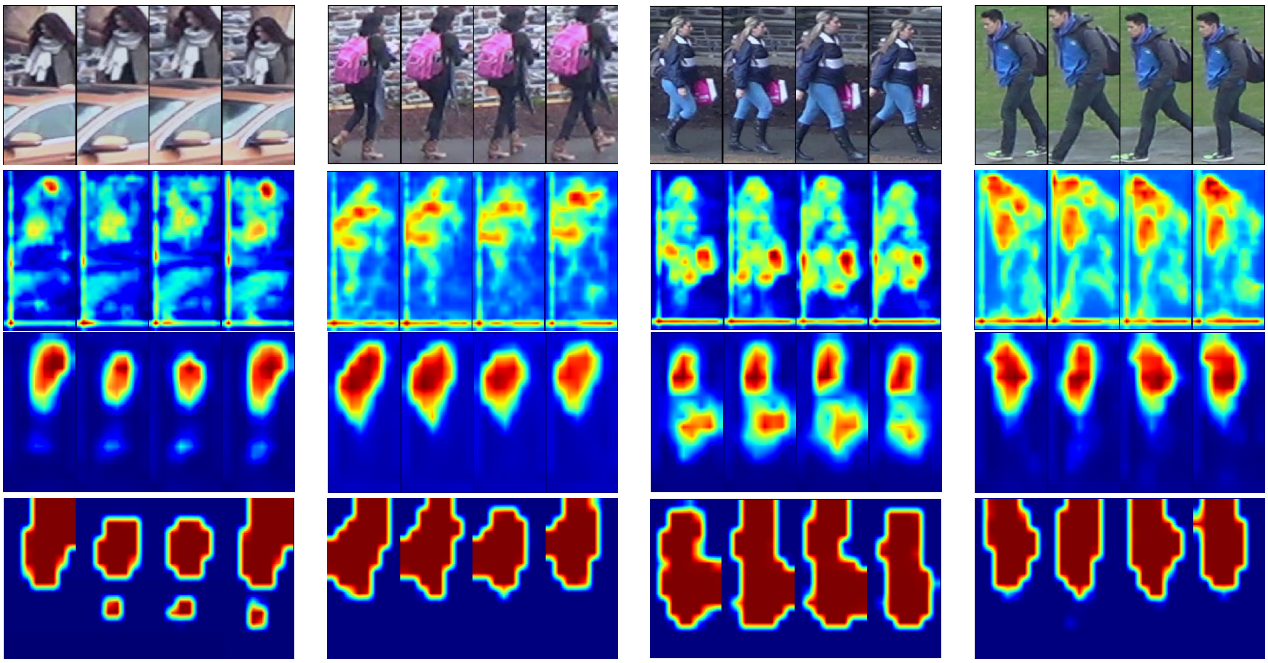
We observe that SE-ResNet50 (COSAM$_{4,5}$) along with TP$_{avg}$ aggregation layer outperform the nearest competing method [2] by 2.5% mAP CMC Rank-1 (Table 5).

Method	iLIDS-VID		
	R1	R5	R20
Top push video Re-ID [4]	56.3	87.6	98.3
JST-RNN[5]	55.2	86.5	97.0
Joint ST pooling [3]	62.0	86.0	98.0
Region QEN[2]	77.1	93.2	99.4
RevisitTempPool[1]	73.9	92.6	98.41
[1] + SE-ResNet50 + TP$_{avg}$	76.87	93.94	99.07
SE-ResNet50 + COSAM$_{4,5}$ + TP$_{avg}$(ours)	**79.61**	**95.32**	**99.8**

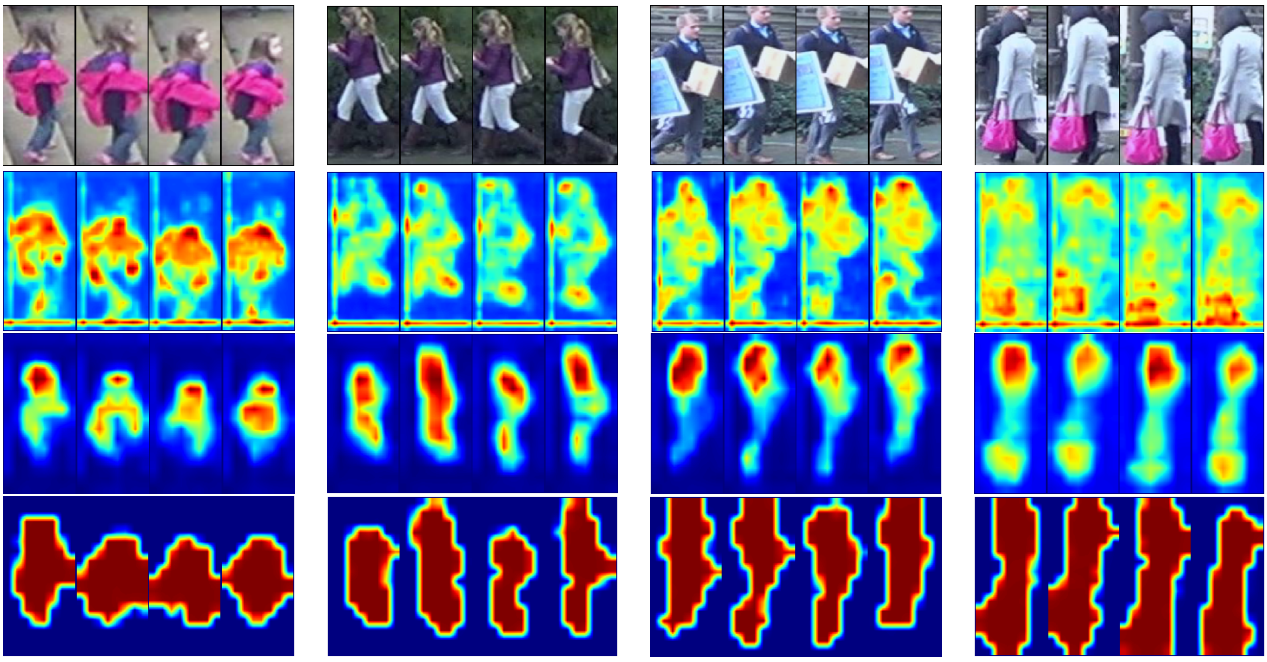
Table 5. Comparison of the state-of-the-arts in iLIDS-VID dataset.

4. Visualization of attention masks

We visualize the spatial masks obtained from the test sets of DukeMTMC-VideoReID and MARS in Figures 1, 2, 3, 4.



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



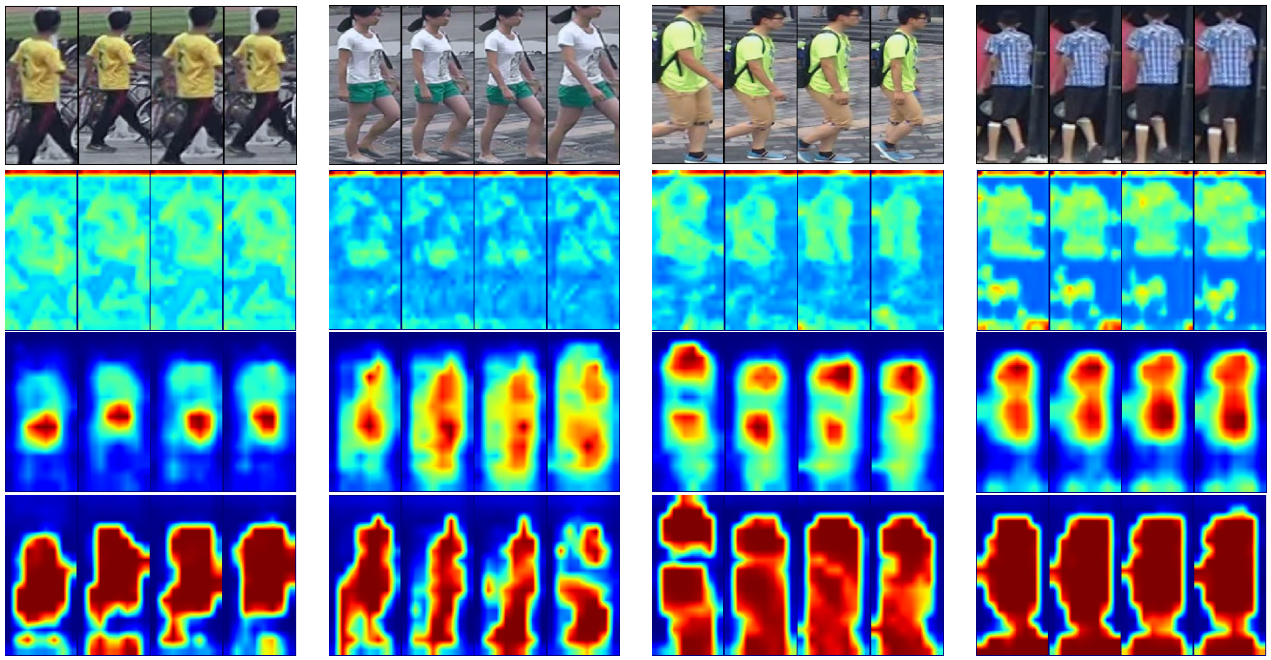
(e) (f) (g) (h)

Figure 1. Visualization results. In each column, the first row shows the image frames of a person from **DukeMTMC-VideoReID dataset**. The second, third and fourth rows show the corresponding co-segmentation maps at COSAM_3, COSAM_4 and COSAM_5 layers respectively, in SE-ResNet50+$\text{COSAM}_{3,4,5}$ model trained on DukeMTMC-VideoReID dataset.

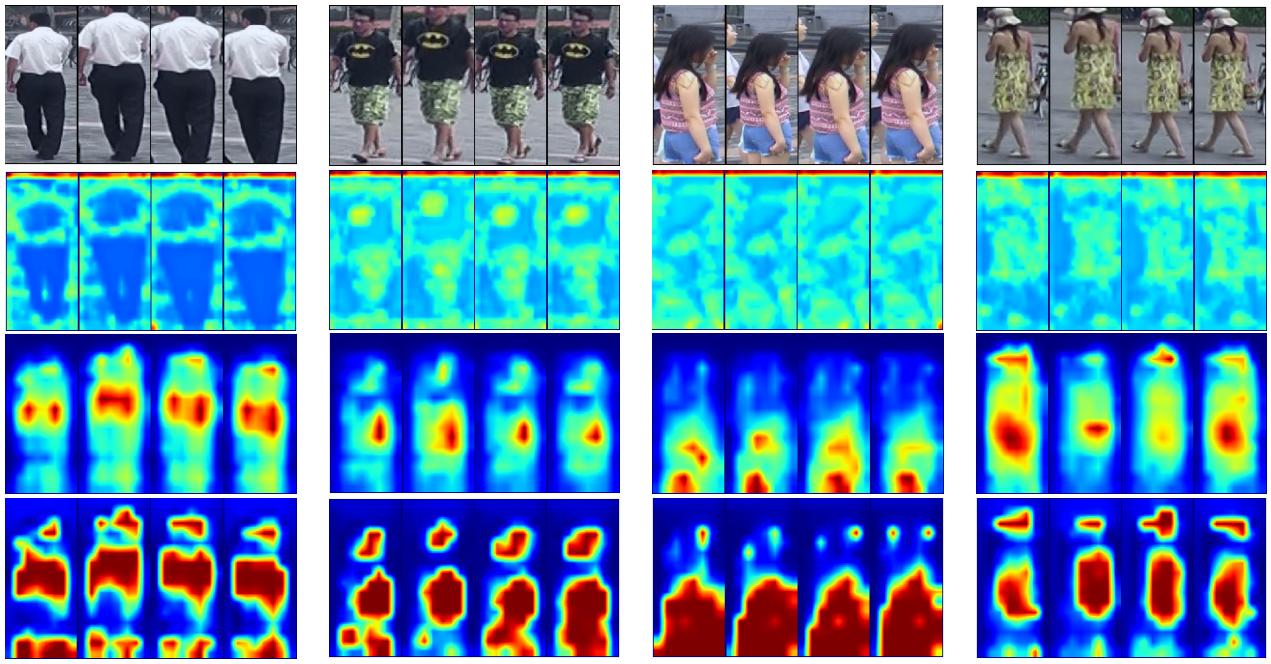
References

[1] J. Gao and R. Nevatia. Revisiting temporal modeling for video-based person reid. *CoRR*, abs/1805.02104, 2018. 2

[2] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai. Region-based quality estimation network for large-scale person re-identification. In *Thirty-Second AAAI Conference on Artificial*



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

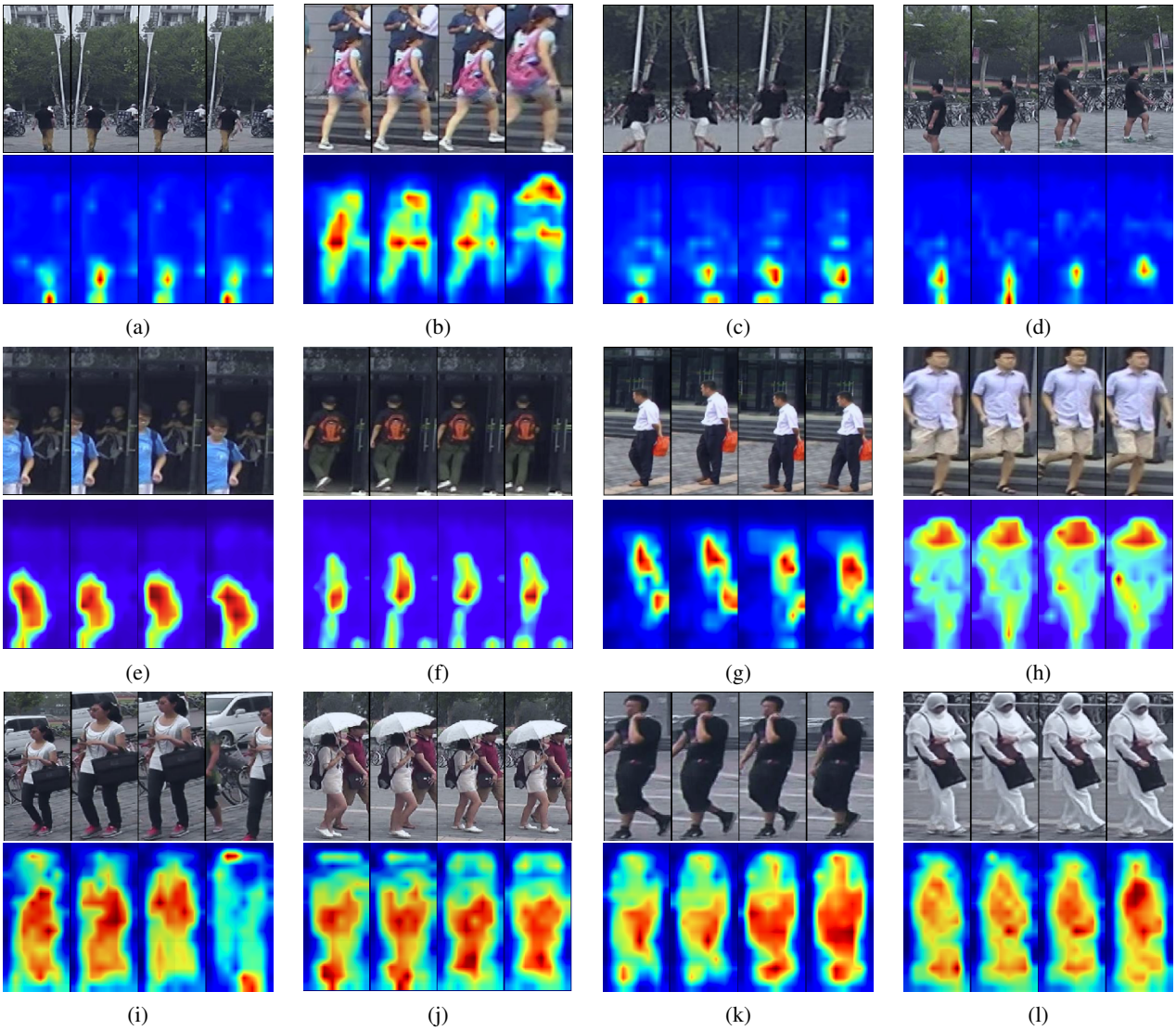


(e) (f) (g) (h)

Figure 2. Visualization results. In each column, the first row shows the image frames of a person from **MARS dataset**. The second, third and fourth rows show the corresponding co-segmentation maps at COSAM_3, COSAM_4 and COSAM_5 layers respectively, in SE-ResNet50+$\text{COSAM}_{3,4,5}$ model trained on MARS dataset.

Intelligence, 2018. 2

[3] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4733–4742, 2017. 2

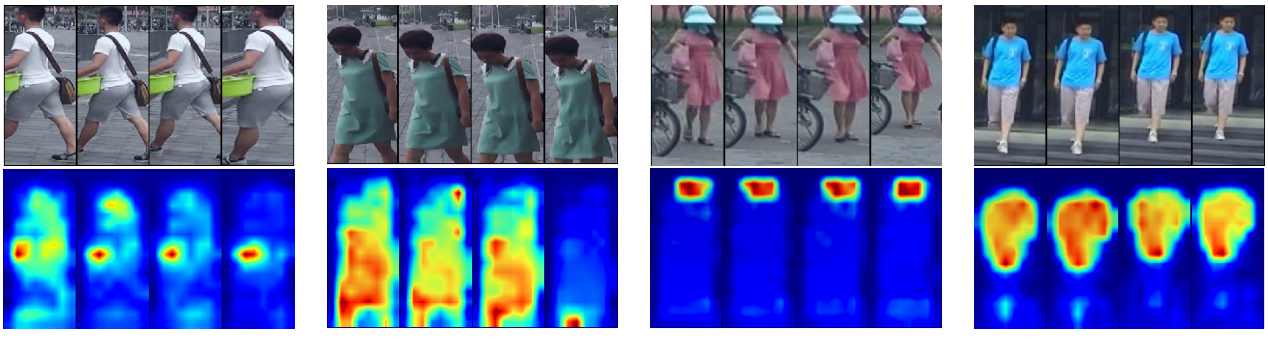


(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 3. Visualization results. The second row shows the co-segmentation maps (COSAM_4 layer in SE-ResNet50+$\text{COSAM}_{4,5}$ model trained on MARS dataset) corresponding to the images in the first row.



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

Figure 4. Visualization of some of the failure cases in the COSAM spatial attention step disregarding certain part informations. The second row shows the co-segmentation maps (COSAM_4 layer in SE-ResNet50+$\text{COSAM}_{4,5}$ model trained on MARS dataset) corresponding to the images in the first row.

[4] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2016. 2

[5] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017. 2