Co-segmentation Inspired Attention Networks for Video-based Person Re-identification

Arulkumar Subramaniam, Athira Nambiar and Anurag Mittal

Indian Institute of Technology Madras

ICCV 2019 Seoul, Korea

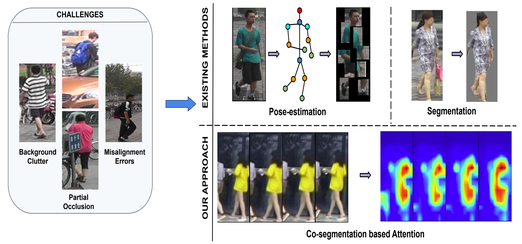
Objectives

Problem definition : Person Re-identification is the task of identifying a particular person across multiple images captured from the same/different cameras, from different viewpoints, at the same/different points in time.

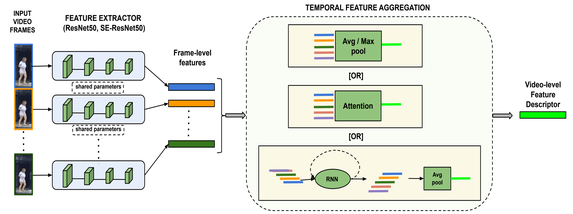
Contributions :

1. A novel **Co-segmentation inspired Re-ID** architecture.
2. **"Co-segmentation Activation Module (COSAM)"** that can be included in any deep neural architecture to enhance common abstract features and to suppress background features.
3. Co-segmentation masks pave way towards "**Interpretable**" deep learning.

Goal: Enhanced Object Attention in Re-ID



A Standard Video-based Re-ID pipeline [1]



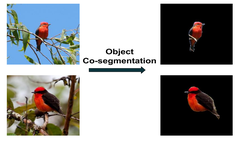
Baseline performances

Feature extractor	Temporal aggregation	MARS				DukeMTMC-VideoReID			
		mAP	R1	R5	R20	mAP	R1	R5	R20
ResNet50[1]	Temporal Average Pooling (TP_{avg})	75.8	83.1	92.8	96.8	92.9	93.6	**99.0**	**99.7**
ResNet50[1]	Temporal Attention (TA)	76.7	83.3	93.8	**97.4**	93.2	93.9	98.9	99.5
ResNet50[1]	Recurrent Neural Network (RNN)	73.8	81.6	92.8	96.7	88.1	88.7	97.6	99.3
SE-ResNet50	Temporal Average Pooling (TP_{avg})	**78.1**	84.0	**95.2**	97.1	**93.5**	93.7	**99.0**	**99.7**
SE-ResNet50	Temporal Attention (TA)	77.7	**84.2**	94.7	**97.4**	93.1	**94.2**	**99.0**	**99.7**
SE-ResNet50	Recurrent Neural Network (RNN)	75.7	83.1	93.6	96.0	92.4	94.0	98.4	99.1

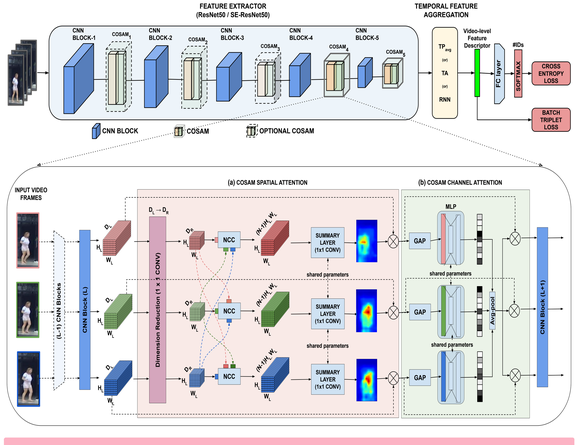
Code: `https://github.com/InnovArul/vidreid_cosegmentation`

Inspiration from Object Co-segmentation task

1. Task of identifying and segmenting common objects from two or more images (Li *et al.*[2] & Chen *et al.*[3]).



Proposed Network Architecture



COSAM components

1. **Dimension reduction step** to reduce the number of channels from D_L to D_R ($D_R << D_L$).
2. **Spatial Attention step** estimates a spatial mask for each frame via mutual consensus across frames, carried out using **Normalized Cross Correlation (NCC)**.

   $F_{n,p}$ = Feature map of n^{th} frame of person p (I_n^p) of dimension $D_R \times H_L \times W_L$ (D_R = Number of channels, H_L = Height, W_L = Width).

   $F_{n,p}^{(i,j)} = D_R$ dimensional local descriptor of the frame at location (i, j).

$$\text{Cost Volume}(i,j) = \{NCC\left(F_{n,p}^{(i,j)}, F_{m,p}^{(h,w)}\right) \mid (1 \le m \le N, m \ne n) \;\&\; (1 \le h \le H_L) \;\&\; (1 \le w \le W_L)\}$$

$$\text{NCC}(P,Q) = \frac{1}{D_R} \frac{\sum_{k=1}^{D_R} (P_k - \mu_P).(Q_k - \mu_Q)}{\sigma_P . \sigma_Q}$$

   Here, (μ_P, μ_Q) denote the mean of the descriptors (P, Q), & (σ_P, σ_Q) denote the standard deviations of the descriptors (P, Q) respectively.

3. **Channel Attention step** determines the importance of each channel by consulting with all the frames.

Objective functions

$$L = \sum_{i=1}^{B} \{L_{CE} + \lambda L_{triplet}(I_i, I_{i_+}, I_{i_-})\} \quad (1)$$

1. **Cross-Entropy loss (L_{CE}):** Supervised loss for person classification
2. **Batch triplet loss ($L_{triplet}$):** To reduce the intra-class variation and to increase the inter-class variation.

$$L_{triplet}(I_A, I_+, I_-) = \max\{D(f_{I_A}, f_{I_+}) - D(f_{I_A}, f_{I_-}) + m, 0\} \quad (2)$$

I_A, I_+, I_- are the anchor, positive and negative examples respectively & m = margin.

Ablation study

	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**
	$COSAM_{4,5}$	**77.2**	**83.7**	94.1	**97.5**	**94.0**	94.4	**99.1**	**99.9**
SE-ResNet50	No COSAM	78.1	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**
	$COSAM_{4,5}$	**79.9**	84.9	95.5	**97.9**	**94.1**	**95.4**	**99.3**	99.8

Table: Study of COSAM's performance at different positions with temporal aggregation as TP_{avg}.

	Temp. Agg.	COSAM$_i$	MARS			Duke			iLIDS-VID	
			mAP	R1	R5	mAP	R1	R5	R1	R5
ResNet50	TP_{avg}[1]	-	75.8	83.1	92.8	92.9	93.6	99.0	73.9	92.6
	TP_{avg}	$COSAM_{4,5}$	**77.2**	**83.7**	94.1	**94.0**	94.4	99.1	**75.5**	94.1
	TA [1]	-	76.7	83.3	93.8	93.2	93.9	98.9	72.3	92.4
	TA	$COSAM_{4,5}$	76.9	83.6	93.7	93.4	**94.6**	98.9	74.9	94.4
	RNN[1]	-	73.8	81.6	92.8	88.1	88.7	97.6	68.5	93.2
	RNN	$COSAM_{4,5}$	74.8	82.4	93.9	90.4	91.7	98.3	68.9	93.1
SE-ResNet50	TP_{avg}	-	78.1	84.0	95.2	93.5	93.7	99.0	76.9	93.9
	TP_{avg}	$COSAM_{4,5}$	**79.9**	84.9	95.5	**94.1**	**95.4**	99.3	**79.6**	95.3
	TA	-	77.7	84.2	94.7	93.1	94.2	99.0	74.7	93.2
	TA	$COSAM_{4,5}$	79.1	**85.0**	94.9	94.1	**95.3**	98.9	77.1	94.7
	RNN	-	75.7	83.1	93.6	92.4	94.0	98.4	77.4	94.4
	RNN	$COSAM_{4,5}$	76.0	83.4	93.9	92.5	93.9	98.3	77.8	97.3

Table: Best mAP & CMC Rank-1 per backbone network are shown in **red** and **blue** respectively.

Results

Network	Deep model?	MARS			
		mAP	R1	R5	R20
LOMO+XQDA	No	16.4	30.7	46.6	60.9
IDE+XQDA+ReRanking	Yes	68.5	73.9	-	-
Region QEN	Yes	71.1	77.8	88.8	94.1
Comp. Snippet Sim.	Yes	69.4	81.2	92.1	-
Part-Aligned	Yes	72.2	83.0	92.8	96.8
RevisitTempPool[1]	Yes	76.7	83.3	93.8	97.4
[1] + SE-ResNet50 + TP_{avg}	Yes	78.1	84.0	95.2	97.1
SE-ResNet50 + $COSAM_{4,5}$ + TP_{avg}(ours)	Yes	**79.9**	**84.9**	**95.5**	**97.9**
SE-ResNet50 + $COSAM_{4,5}$ + TP_{avg}(ours) + Re-ranking	Yes	**87.4**	**86.9**	**95.5**	**98.0**
Network	Deep model?	DukeMTMC-VideoReID			
		mAP	R1	R5	R20
ETAP-Net	Yes	78.34	83.62	94.59	97.58
RevisitTempPool[1]	Yes	93.2	93.9	98.9	99.5
[1] + SE-ResNet50 + TP_{avg}	Yes	93.5	93.7	99.0	99.7
SE-ResNet50 + $COSAM_{4,5}$ + TP_{avg}(ours)	Yes	**94.1**	**95.4**	**99.3**	**99.8**

Table: State-of-the-art comparison

Model	Handbag			Hat			Backpack		
	mAP	R1	R5	mAP	R1	R5	mAP	R1	R5
[1]	91.2	92.0	**100.0**	91.1	91.7	97.5	92.8	93.9	98.6
R50+$C_{4,5}$	**95.2**	**96.0**	**100.0**	**93.5**	**94.2**	**97.5**	**95.1**	**96.4**	**99.8**
[1]+SE50	94.1	97.3	**100.0**	92.7	94.2	99.2	94.3	95.6	99.1
SE50+$C_{4,5}$	**96.0**	**100.0**	**100.0**	**93.9**	**96.7**	**99.5**	**95.4**	**97.1**	**100.0**

Table: Attribute-wise performance improvements on Duke. Here R50 = ResNet50, SE50 = SE-ResNet50 & $C_{4,5}$ = $COSAM_{4,5}$.

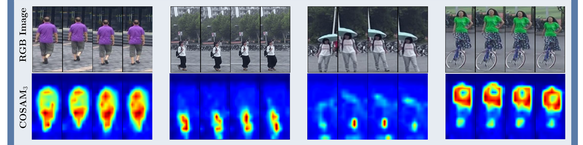
frame length	MARS				DukeMTMC-VideoReID			
	mAP	R1	R5	R20	mAP	R1	R5	R20
$N = 2$	78.1	83.5	94.3	98.1	94.0	94.3	99.1	**99.9**
$N = 4$	**79.9**	**84.9**	**95.5**	**97.9**	**94.1**	**95.4**	**99.3**	99.8
$N = 8$	77.4	84.6	94.2	97.0	92.1	91.9	99.0	99.6

Table: Influence of Number of frames (N)

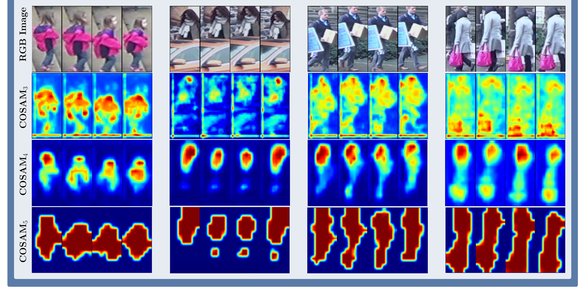
frame selection	MARS				DukeMTMC-VideoReID			
	mAP	R1	R5	R20	mAP	R1	R5	R20
sequential	**79.9**	**84.9**	**95.5**	**97.9**	**94.1**	**95.4**	**99.3**	**99.8**
random	77.5	83.3	93.6	97.0	90.2	90.6	98.3	99.6

Table: Influence of frame selection

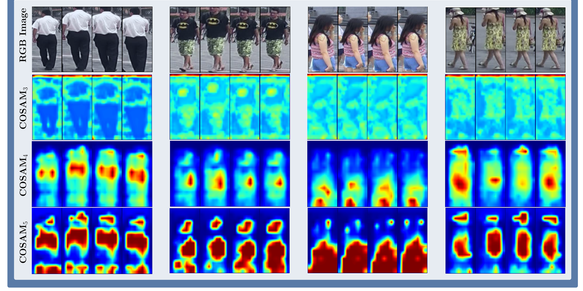
Visualization of the Co-segmentation masks



Co-segmentation masks in DukeMTMC-VideoReID Testset



Co-segmentation masks in MARS Testset



References

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

[2] W. Li, O. H. Jafari, and C. Rother, "Deep object co-segmentation," *arXiv preprint arXiv:1804.06423*, 2018.

[3] H. Chen, Y. Huang, and H. Nakayama, "Semantic aware attention based deep object co-segmentation," *arXiv preprint arXiv:1810.06859*, 2018.

Acknowledgements

This work is supported by grants from Prime Minister's fellowship for Doctoral Research (SERB, India) & Google PhD Fellowship to Arulkumar Subramaniam.