# Co-segmentation Inspired Attention Module for Video-based Vision tasks

Arulkumar S

under the guidance of Prof.Anurag Mittal

Computer Vision Lab, IIT Madras

Seminar-II

# Outline

# Person Re-Identification
## Problem definition

- A fine-grained retrieval task to match a person's image with images from database
- images captured at
  - same/different points in time (of same day)
  - same/different camera
  - various lighting conditions + unconstrained viewpoint/pose changes
- No information about camera position, intrinsic and extrinsic parameters

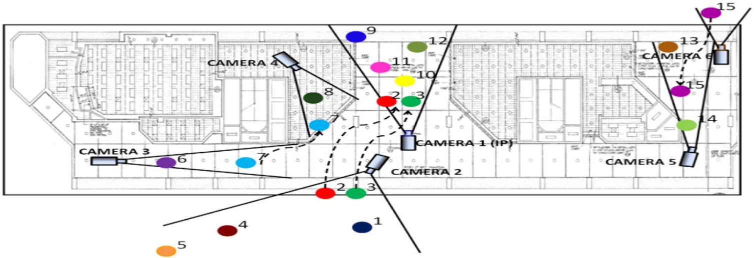
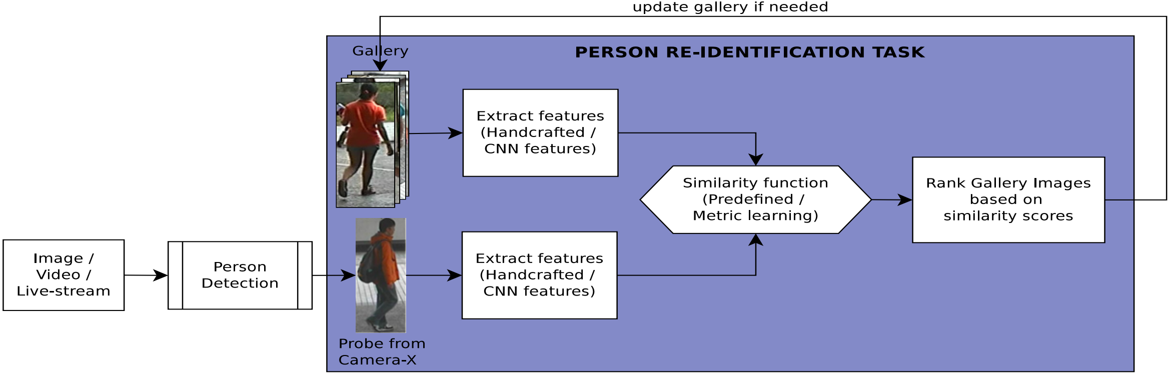



**Image from:** Apurva *et al.* **A survey of approaches and trends in person re-identification**. Image and Vision Computing - 2014.

## Person Re-Identification setup

- **Probe:** the person's image(s) to be searched in the database
- **Gallery:** one (or more) unique image(s) of persons observed so far. Usually, Gallery images will be available in a database.



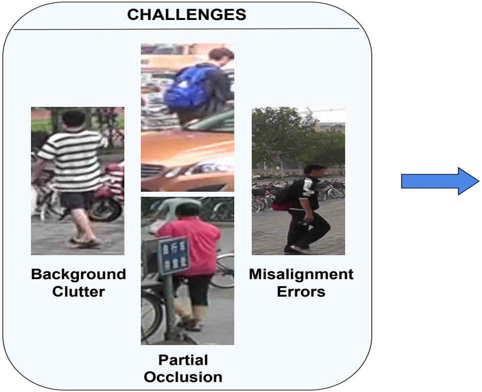

- **Evaluation:** Ranking of matching scores (rank-1, rank-5, . . . ), mean average precision (mAP)

## Practical challenges

- Illumination variation
- Pose/Viewpoint variation
- Background clutters / misalignment errors
- Partial occlusion
- Bad quality images

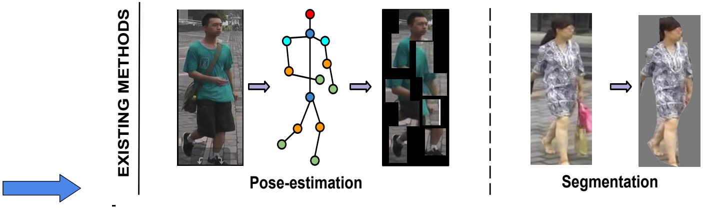
Viewpoint Change | Illumination Variation | Partial Occlusion | Poor quality of images

## Seminar-II: Background clutter / Misalignment errors

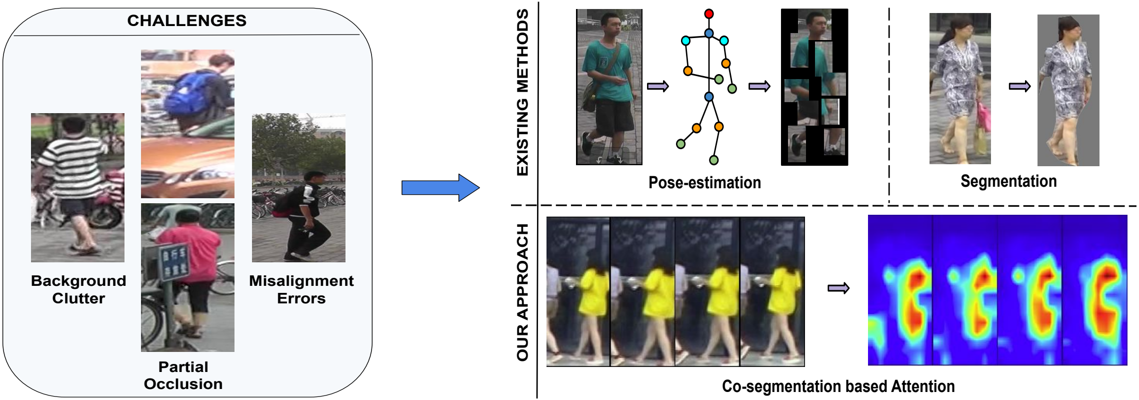


## Seminar-II: Background clutter / Misalignment errors

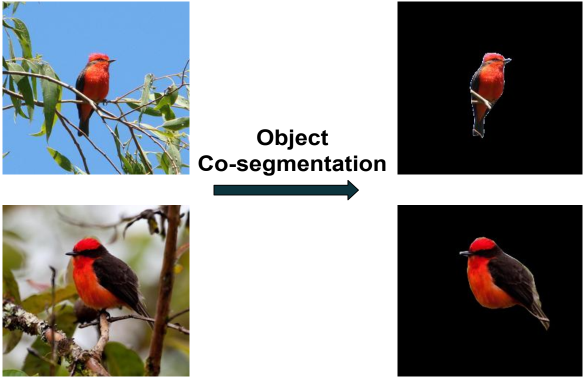
## Seminar-II: Background clutter / Misalignment errors

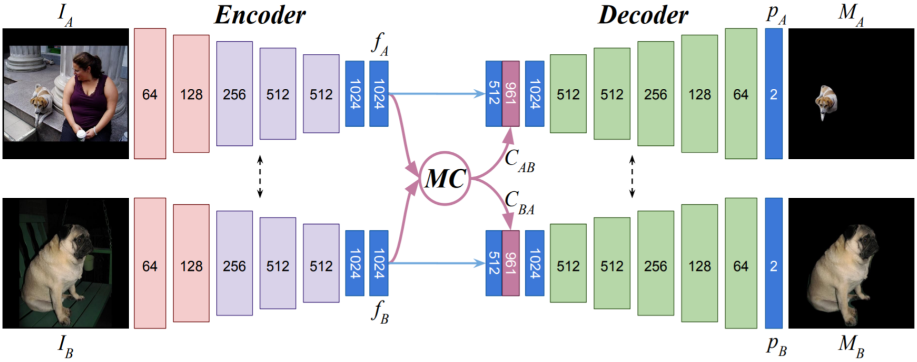
## Seminar-II: Background clutter / Misalignment errors



## Co-segmentation concept



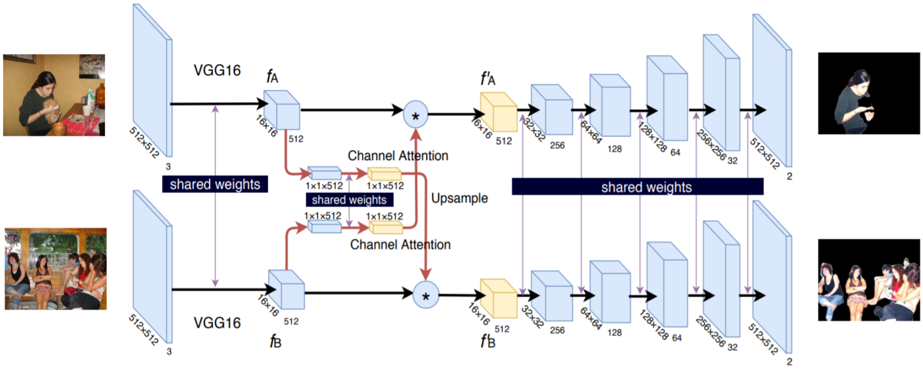
## Co-segmentation in Deep learning literature




*MC = Mutual Correlation

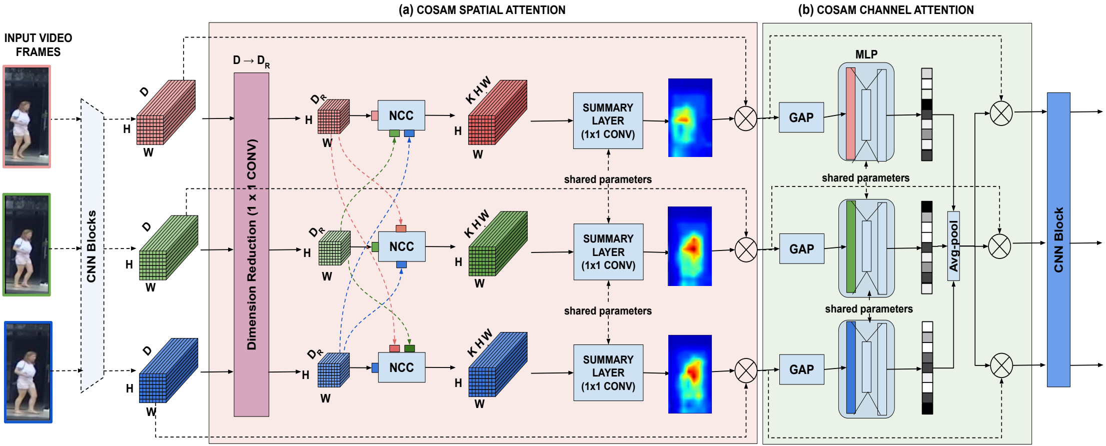
Weihao Li, Omid Hosseini Jafari, and Carsten Rother. "Deep object co-segmentation." Asian Conference on Computer Vision. Springer, Cham, 2018.

## Co-segmentation in Deep learning literature

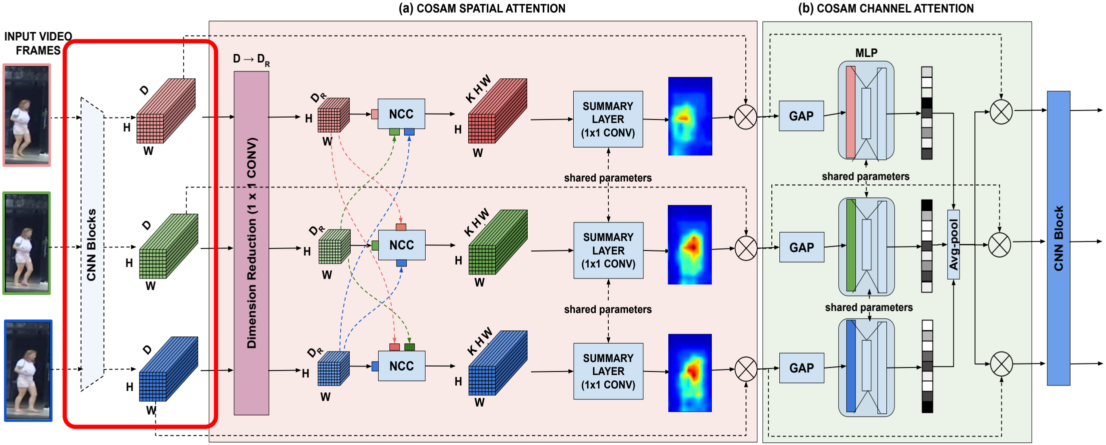



Hong Chen, Yifei Huang, and Hideki Nakayama. "Semantic aware attention based deep object co-segmentation." Asian Conference on Computer Vision. Springer, Cham, 2018.

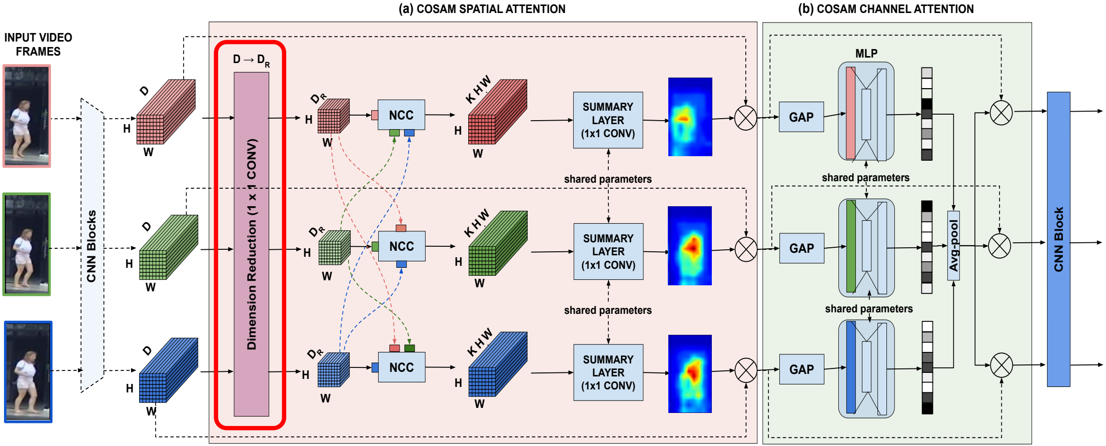
# Co-segmentation Activation Module (COSAM)




Input $(N \times D \times H \times W) \rightarrow$ Induce co-segmentation $\rightarrow$ Output $(N \times D \times H \times W)$

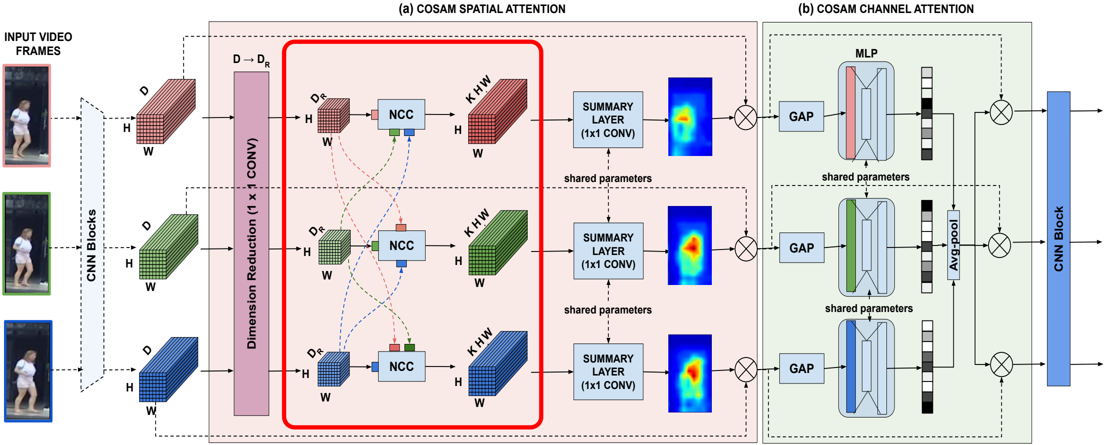



Frames of dimension $N \times 3 \times H_I \times W_I$ are passed through $L$ CNN blocks to get feature maps of dimension $N \times D \times H \times W$.



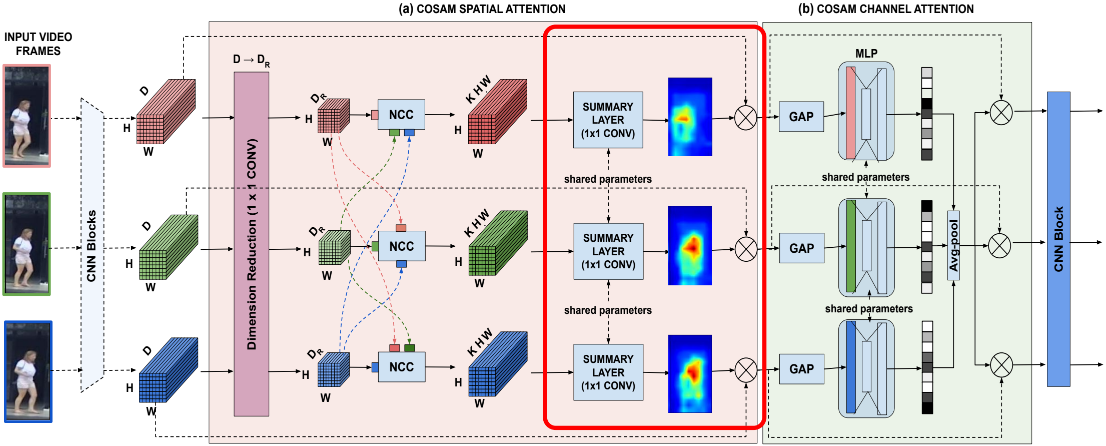

Dimensionality reduction ($D \times H \times W \longrightarrow D_R \times H \times W$)

$D \longrightarrow D_R$ - to reduce computational overhead

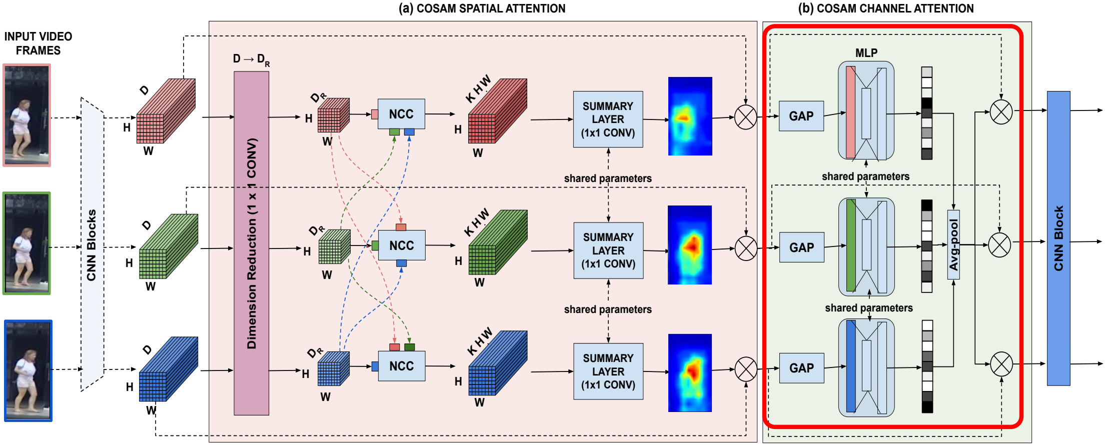



$$\text{Cost volume}_{(n)}(i,j) = \{NCC\left(F_n^{(i,j)}, R_k^{(h,w)}\right) \quad 1 \leq k \leq K, 1 \leq h \leq H, 1 \leq w \leq W\} \tag{1}$$

$$\text{NCC}(P,Q) = \frac{1}{D_R}\frac{\sum_{k=1}^{D_R}(P_k - \mu_P).(Q_k - \mu_Q)}{\sigma_P.\sigma_Q} \tag{2}$$

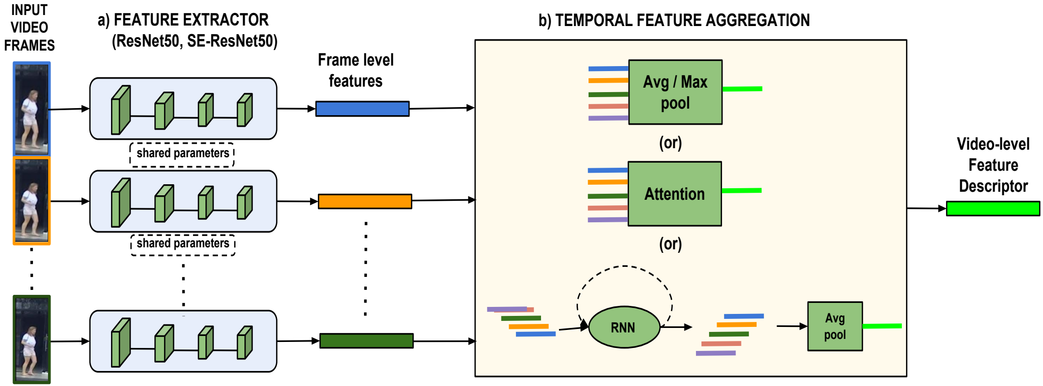



- Pass cost volume through Conv + BN + ReLU → Sigmoid to get spatial mask.
- Multiply spatial masks with corresponding feature maps




- Per-frame Channel attention from Global Average Pool-ed (GAP) feature maps
- Average of per-frame channel attentions to capture common important channels

## Video Re-ID pipeline

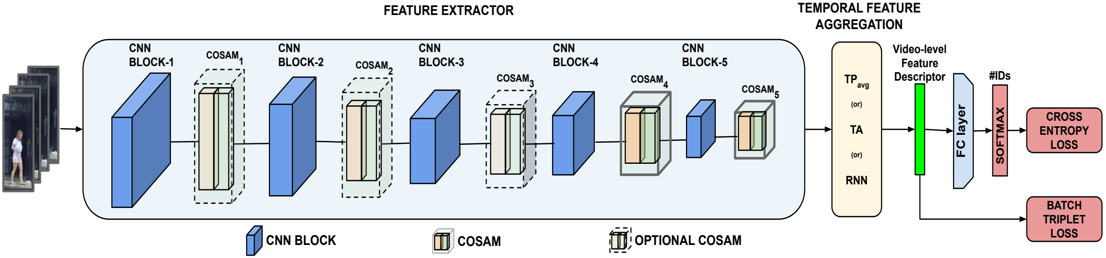



Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. **Co-segmentation Inspired Attention Networks for Video-based Person Re-identification**. Proceedings of the International Conference on Computer Vision (ICCV) - 2019.

## Video Re-ID datasets

- MARS
  - 1261 identities and 20,478 video sequences
  - 6 non-overlapping cameras
  - 625 identities for training and the rest for testing
  - Additional 3,248 identities for distractors
- DukeMTMC-VideoReID
  - 702 identities each for training and testing
  - 369,656 tracklets for training, and 445,764 frames for testing
  - 402 identities for distractors
- iLIDs-VID
  - Small dataset
  - 300 persons each for training and testing

## Model architecture




Training loss function:

$$L = \sum_{i=1}^{B} \left\{ L_{CE} + \lambda L_{triplet}(I_i, I_{i_+}, I_{i_-}) \right\} \tag{3}$$

## COSAM at different levels

| | $COSAM_i$ | 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: Evaluation of the backbone feature extractors with COSAM plugging in after $i^{th}$ *CNN* block.

## COSAM at different levels

| | $COSAM_i$ | 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: Evaluation of the backbone feature extractors with COSAM plugging in after $i^{th}$ *CNN* block.

## COSAM with different temporal modeling schemes

| | 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: Comparison of the baseline models with best performing COSAM-configuration ($COSAM_{4,5}$). Best mAP & CMC Rank-1 per backbone network are shown in **red** and **blue** colors respectively.

[1] Jiyang Gao, and Ram Nevatia. "Revisiting temporal modeling for video-based person reid." arXiv preprint arXiv:1805.02104 (2018).

## Comparison with State-of-the-arts

| Network | Deep model? | MARS | | | |
|---|---|---|---|---|---|
| | | mAP | R1 | R5 | R20 |
| TriNet | Yes | 67.7 | 79.8 | 91.4 | - |
| 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 | Yes | 76.7 | 83.3 | 93.8 | 97.4 |
| 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 | Yes | 93.2 | 93.9 | 98.9 | 99.5 |
| 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** |

## Comparison with State-of-the-arts

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

## Number of reference frames

| 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: Evaluation of the influence of track length $T$ on Re-ID performance in *SE-ResNet50+COSAM*$_{4,5}$+*TP*$_{avg}$.

## Attribute-wise performance

| Model | Handbag | | | Hat | | | Backpack | | |
|---|---|---|---|---|---|---|---|---|---|
| | mAP | R1 | R5 | mAP | R1 | R5 | mAP | R1 | R5 |
| ResNet50+TP | 91.2 | 92.0 | **100.0** | 91.1 | 91.7 | 97.5 | 92.8 | 93.9 | 98.6 |
| ResNet50+COSAM$_{4,5}$+TP | **95.2** | **96.0** | **100.0** | **93.5** | **94.2** | **97.5** | **95.1** | **96.4** | **99.8** |
| SE-ResNet50+TP | 94.1 | 97.3 | **100.0** | 92.7 | 94.2 | 99.2 | 94.3 | 95.6 | 99.1 |
| SE-ResNet50+COSAM$_{4,5}$+TP | **96.0** | **100.0** | **100.0** | **93.9** | **96.7** | **99.5** | **95.4** | **97.1** | **100.0** |

Table: Attribute-wise performance comparison on Duke dataset. TP = Temporal average pooling.

## Cross-dataset performance

| | 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** |

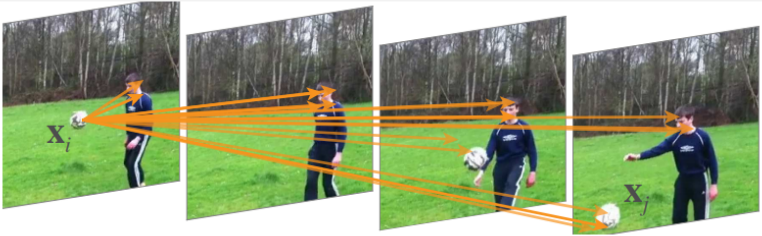
Table: Cross-dataset performance of the best performing model with *SE-ResNet50* as the feature extractor and $TP_{avg}$ as the temporal aggregation layer. Here *DukeMTMC* = DukeMTMC-VideoReID.

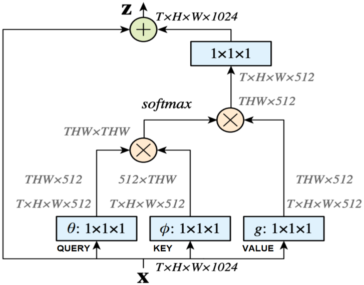
## Spatial vs. Channel attention

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

Table: 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*$_{avg}$.

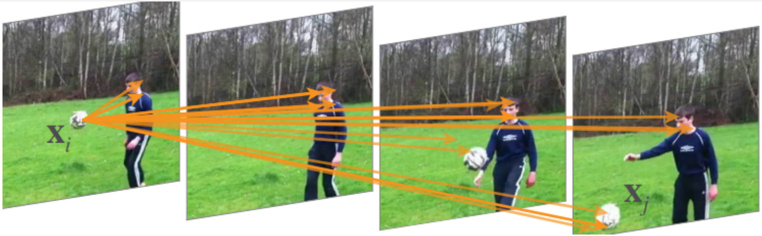
## COSAM vs. Non-local Module (NLM)

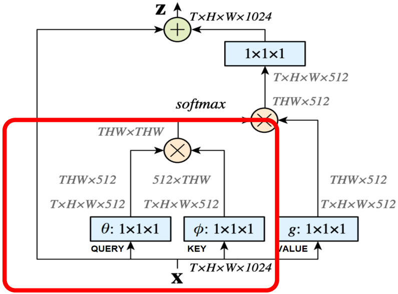



Xiaolong Wang, et al. "Non-local neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

## COSAM vs. Non-local Module (NLM)




Xiaolong Wang, et al. "Non-local neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

## COSAM vs. Non-local Module (NLM)

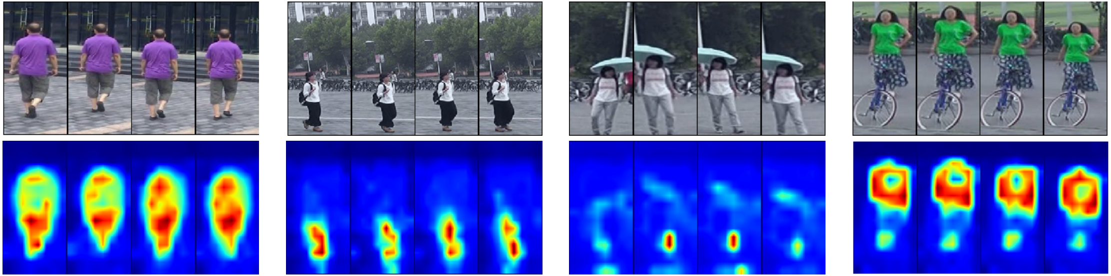
| Module | | #Params | #FLOPs |
|---|---|---|---|
| **NLM** | Gauss. | 4.2M | 4.3B |
| | Gaussian embedding | 8.39M | 8.59G |
| | Concatenation | 8.4M | 8.72G |
| | Dot product | 8.39M | 8.59G |
| **COSAM (ours)** | | **1.6M** | **0.57G** |

Table: COSAM *vs.* Non-local Module (input = $4 \times 2048 \times 16 \times 8$).
Observation: COSAM uses $\sim 4x$ less memory and $\sim 16x$ less computation than NLM.

| Model | #Params | #FLOPs | MARS | | |
|---|---|---|---|---|---|
| | | | mAP | R1 | R5 |
| ResNet50+NLM$_{4,5}$+TP | 34.31M | 27.11B | 76.9 | 83.2 | **94.2** |
| ResNet50+COSAM$_{4,5}$+TP | 26.22M | 17.24B | **77.2** | **83.7** | 94.1 |
| SE-ResNet50+NLM$_{4,5}$+TP | 36.85M | 26.74B | 77.9 | 83.3 | 94.7 |
| SE-ResNet50+COSAM$_{4,5}$+TP | 28.76M | 16.86B | **79.9** | **84.9** | **95.5** |

Table: Comparison of COSAM *vs.* Non-local Module on MARS dataset.

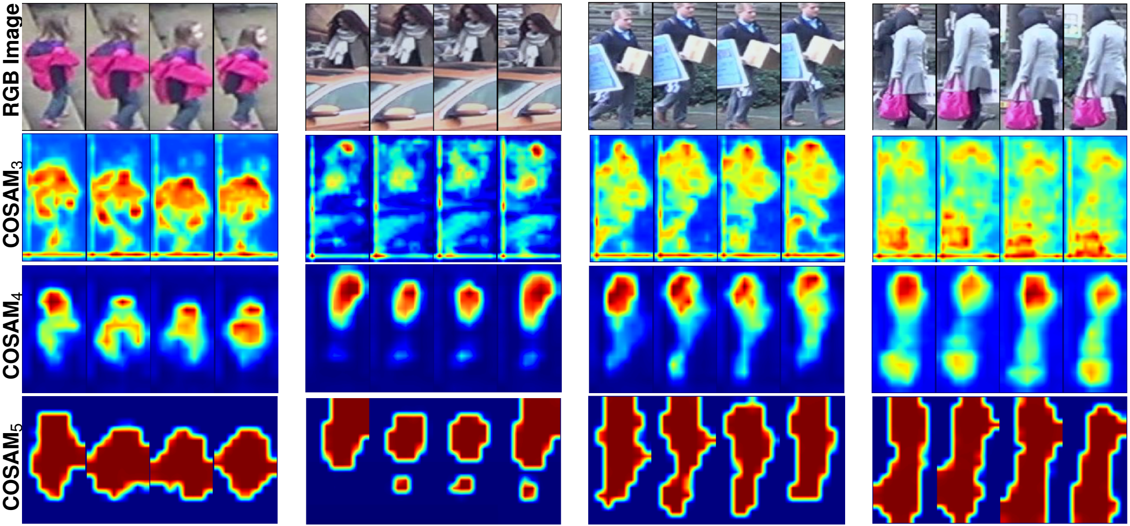
## Qualitative visualization



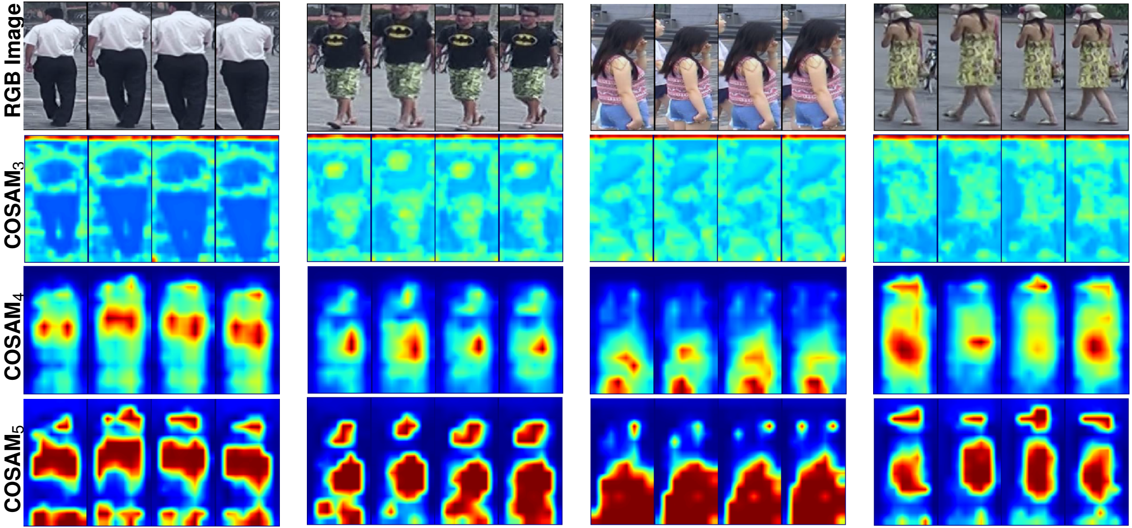
(a) (b) (c) (d)

Figure: Visualization of co-segmentations. The second row shows the segmentation maps corresponding to the images in the first row.

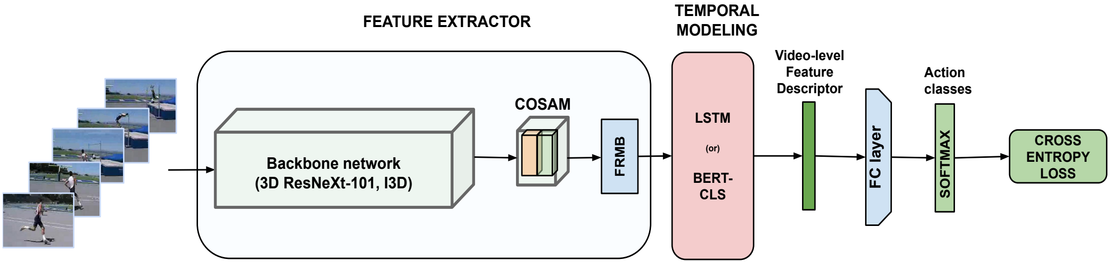
## Qualitative visualization



## Qualitative visualization



## Extending to Video classification task




*FRMB = Feature Reduction with Modified Block

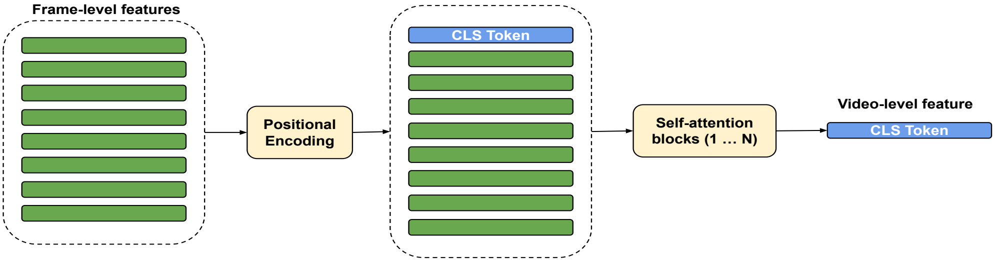
Training via simple cross-entropy loss:

$$L = -\sum_{k=1}^{N} \sum_{c=1}^{C} I(c = t_k) \log p_k^c \tag{4}$$

Here, $I(.)$ denotes an indicator function, $C$ = number of classes, $N$ = number of videos, $t_k$ = the target class one-hot vector, class softmax probabilities $\{p_k^j\}_{j=1}^{C}$.

[2] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition." European Conference on Computer Vision. Springer, Cham, 2020.

## Working of BERT-CLS



## Video classification datasets

- HMDB51
  - 51 action categories
  - total of 6,766 video clips extracted from movie scenes and YouTube.
  - predefined split of train and test sequences
- UCF101
  - Total of 13220 videos belonging to 101 action classes
  - average length of 180 frames per video
  - predefined split of train and test sequences

## Quantitative results

| Backbone | COSAM? | temporal modeling? | #params (M) | #Flops (G) | HMDB51 | | UCF101 | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Top-1% | Top-3% | Top-1% | Top-3% |
| ResNeXt101 [2] | ✗ | LSTM | 47.6 | 38.64 | 73.68 | 87.46 | 93.90 | 98.05 |
| ResNeXt101 | ✓ | LSTM | 48.41 | 38.77 | **75.16** | **89.22** | **94.59** | **98.52** |
| ResNeXt101 [2] | ✗ | BERT | 47.4 | 38.37 | 76.08 | 90.46 | 95.50 | 98.23 |
| ResNeXt101 | ✓ | BERT | 48.21 | 38.49 | **77.52** | **92.55** | **95.96** | **98.84** |
| I3D [2] | ✗ | BERT | 13.57 | 110.6 | 68.63 | 87.78 | 92.50 | 98.26 |
| I3D | ✓ | BERT | 14.23 | 110.7 | 69.38 | 87.95 | 93.05 | 98.63 |

Table: The performance comparison of single stream RGB model from [2] with and without COSAM layer.

[2] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition." European Conference on Computer Vision. Springer, Cham, 2020.

## Quantitative results

| Backbone | COSAM? | temporal modeling? | #params (M) | #Flops (G) | HMDB51 | | UCF101 | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Top-1% | Top-3% | Top-1% | Top-3% |
| ResNeXt101 [2] | ✗ | LSTM | 47.6 | 38.64 | 73.68 | 87.46 | 93.90 | 98.05 |
| ResNeXt101 | ✓ | LSTM | 48.41 | 38.77 | **75.16** | **89.22** | **94.59** | **98.52** |
| ResNeXt101 [2] | ✗ | BERT | 47.4 | 38.37 | 76.08 | 90.46 | 95.50 | 98.23 |
| ResNeXt101 | ✓ | BERT | 48.21 | 38.49 | **77.52** | **92.55** | **95.96** | **98.84** |
| I3D [2] | ✗ | BERT | 13.57 | 110.6 | 68.63 | 87.78 | 92.50 | 98.26 |
| I3D | ✓ | BERT | 14.23 | 110.7 | 69.38 | 87.95 | 93.05 | 98.63 |

Table: The performance comparison of single stream RGB model from [2] with and without COSAM layer.

[2] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition." European Conference on Computer Vision. Springer, Cham, 2020.

## Quantitative results

| Backbone | COSAM? | temporal modeling? | #params (M) | #Flops (G) | HMDB51 | | UCF101 | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Top-1% | Top-3% | Top-1% | Top-3% |
| ResNeXt101 [2] | ✗ | LSTM | 47.6 | 38.64 | 73.68 | 87.46 | 93.90 | 98.05 |
| ResNeXt101 | ✓ | LSTM | 48.41 | 38.77 | **75.16** | **89.22** | **94.59** | **98.52** |
| ResNeXt101 [2] | ✗ | BERT | 47.4 | 38.37 | 76.08 | 90.46 | 95.50 | 98.23 |
| ResNeXt101 | ✓ | BERT | 48.21 | 38.49 | **77.52** | **92.55** | **95.96** | **98.84** |
| I3D [2] | ✗ | BERT | 13.57 | 110.6 | 68.63 | 87.78 | 92.50 | 98.26 |
| I3D | ✓ | BERT | 14.23 | 110.7 | 69.38 | 87.95 | 93.05 | 98.63 |

Table: The performance comparison of single stream RGB model from [2] with and without COSAM layer.

[2] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition." European Conference on Computer Vision. Springer, Cham, 2020.

## Quantitative results

| Backbone | COSAM? | temporal modeling? | #params (M) | #Flops (G) | HMDB51 | | UCF101 | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Top-1% | Top-3% | Top-1% | Top-3% |
| ResNeXt101 [2] | ✗ | LSTM | 47.6 | 38.64 | 73.68 | 87.46 | 93.90 | 98.05 |
| ResNeXt101 | ✓ | LSTM | 48.41 | 38.77 | **75.16** | **89.22** | **94.59** | **98.52** |
| ResNeXt101 [2] | ✗ | BERT | 47.4 | 38.37 | 76.08 | 90.46 | 95.50 | 98.23 |
| ResNeXt101 | ✓ | BERT | 48.21 | 38.49 | **77.52** | **92.55** | **95.96** | **98.84** |
| I3D [2] | ✗ | BERT | 13.57 | 110.6 | 68.63 | 87.78 | 92.50 | 98.26 |
| I3D | ✓ | BERT | 14.23 | 110.7 | 69.38 | 87.95 | 93.05 | 98.63 |

Table: The performance comparison of single stream RGB model from [2] with and without COSAM layer.

[2] M. Esat Kalfaoglu, Sinan Kalkan, and A. Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition." European Conference on Computer Vision. Springer, Cham, 2020.

## State-of-the-art comparisons

| | Method | use flow? | HMDB51 | UCF101 |
|---|---|---|---|---|
| Two-stream | TwoStream | ✓ | 59.40 | 88.00 |
| | TwoStream Fusion + IDT | ✓ | 69.20 | 93.50 |
| | R(2+1)D | ✓ | 78.70 | 97.30 |
| | I3D | ✓ | 80.90 | 97.80 |
| | BubbleNet | ✓ | 82.6 | 97.2 |
| | ResNeXt101 BERT | ✓ | 83.55 | 97.87 |
| Single-stream | IDT | ✗ | 61.70 | - |
| | R(2+1)D | ✗ | 74.50 | 96.80 |
| | MARS + RGB | ✗ | 73.10 | 95.60 |
| | TemporalShift | ✗ | 73.50 | 95.90 |
| | ResNeXt101 BERT | ✗ | 76.08 | 94.59 |
| | **ResNeXt101 + COSAM + BERT (ours)** | ✗ | 77.52 | 95.96 |

Table: State-of-the-art performance comparison of deep models for video action classification task.

## State-of-the-art comparisons

| | Method | use flow? | HMDB51 | UCF101 |
|---|---|---|---|---|
| Two-stream | TwoStream | ✓ | 59.40 | 88.00 |
| | TwoStream Fusion + IDT | ✓ | 69.20 | 93.50 |
| | R(2+1)D | ✓ | 78.70 | 97.30 |
| | I3D | ✓ | 80.90 | 97.80 |
| | BubbleNet | ✓ | 82.6 | 97.2 |
| | ResNeXt101 BERT | ✓ | 83.55 | 97.87 |
| Single-stream | IDT | ✗ | 61.70 | - |
| | R(2+1)D | ✗ | 74.50 | 96.80 |
| | MARS + RGB | ✗ | 73.10 | 95.60 |
| | TemporalShift | ✗ | 73.50 | 95.90 |
| | ResNeXt101 BERT | ✗ | 76.08 | 94.59 |
| | **ResNeXt101 + COSAM + BERT (ours)** | ✗ | 77.52 | 95.96 |

Table: State-of-the-art performance comparison of deep models for video action classification task.

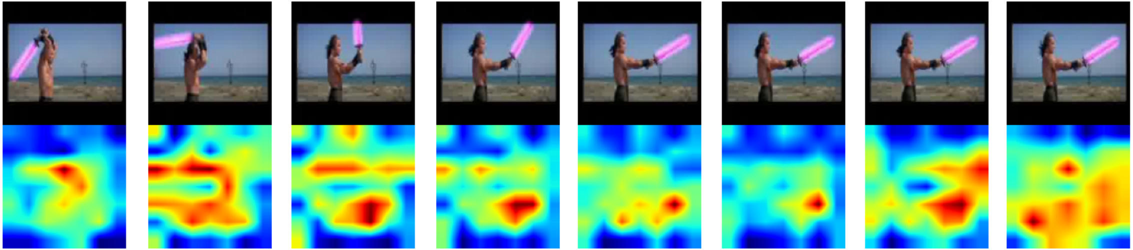


Figure: “Sword Exercise” class from HMDB51 dataset

Figure: “Hit” class from HMDB51 dataset

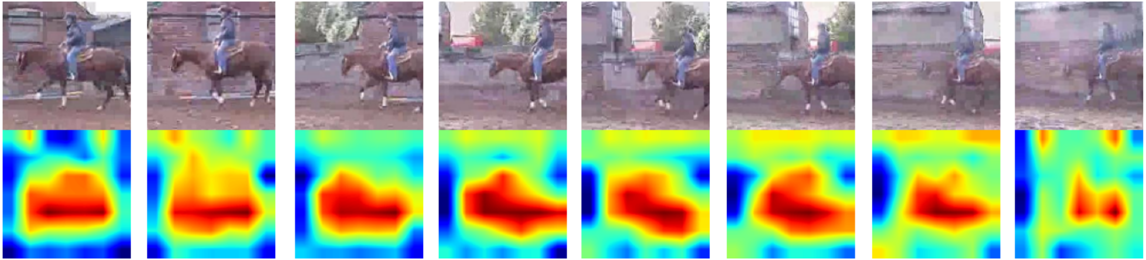


Figure: “Horse Riding” class from UCF101 dataset

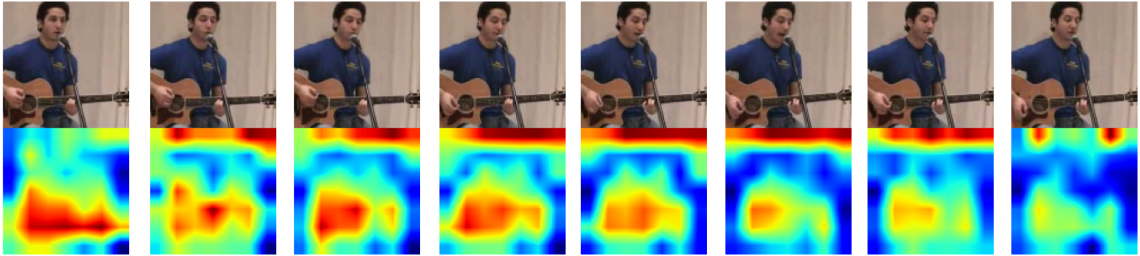


Figure: “Playing Guitar” class from UCF101 dataset

## Summary

- A “co-segmentation” inspired attention module (COSAM) to induce a notion of co-segmentation in feature space.
- COSAM is generic to be applied inside any deep CNN.
- Application to two video based vision tasks:
  - Video based person re-ID
  - Video classification
- **Current work:**
  - Self-supervised contrastive learning for person re-ID

Thank you!

## Journal Articles

Arulkumar Subramaniam, Jayesh Vaidya, Muhammed Abdul Majeed Ameen, Athira Nambiar, and Anurag Mittal. **Co-segmentation Inspired Attention Module for Video-based Computer Vision Tasks**. Submitted to Computer Vision and Image Understanding (CVIU), 2021.

## Conference proceedings

Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. **Co-segmentation Inspired Attention Networks for Video-based Person Re-identification**. Proceedings of the International Conference on Computer Vision (ICCV) - 2019. Seoul, South Korea.

Arulkumar Subramaniam*, Prashanth Balasubramanian*, and Anurag Mittal. **NCC-Net: Normalized Cross Correlation Based Deep Matcher with Robustness to Illumination Variations**. IEEE Winter Conference on the Applications of Computer Vision (WACV) - 2018. Nevada, United States.

Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. **Deep Neural Networks with Inexact Matching for Person Re-Identification**. Proceedings of the Neural Information Processing Systems (NeurIPS) - 2016. Barcelona, Spain.

---

Jayesh Vaidya, Arulkumar Subramaniam, and Anurag Mittal. **Co-Segmentation Aided Two-Stream Architecture for Video Captioning**. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, Hawaii.

Arulkumar Subramaniam*, Ajay Narayanan*, and Anurag Mittal. **Feature Ensemble Networks with Re-ranking for Recognizing Disguised Faces in the Wild**. Proceedings of the International Conference on Computer Vision Workshop (ICCVW) - 2019 on Recognizing Disguised Faces in the Wild.

Arulkumar Subramaniam*, Vismay Patel*, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. **Bi-modal First Impressions Recognition using Temporally Ordered Deep Audio and Stochastic Visual Features**. Proceedings of the European Conference on Computer Vision Workshop (ECCVW) - 2016 on Apparent Personality Analysis. Amsterdam, The Netherlands.