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

Abstract

Person re-identification (Re-ID) is an important real-world surveillance problem that entails associating a person’s identity over a network of cameras. Video-based Re-ID approaches have gained significant attention recently since a video, and not just an image, is often available. In this work, we propose a novel Co-segmentation inspired video Re-ID deep architecture and formulate a Co-segmentation based Attention Module (COSAM) that activates a common set of salient features across multiple frames of a video via mutual consensus in an unsupervised manner. As opposed to most of the prior work, our approach is able to attend to person accessories along with the person. Our plug-and-play and interpretable COSAM module applied on two deep architectures (ResNet50, SE-ResNet50) outperform the state-of-the-art methods on three benchmark datasets.

1. Introduction

Person re-identification (Re-ID) [14] is the task of matching person images/videos across two or more non-overlapping camera views. Recently, it has been drawing significant attention owing to its wide range of applications in surveillance [62], activity analysis [33], etc. However, the problem is challenging due to severe occlusions, background clutter, viewpoint change, etc., and can be thought of as a proxy for other general matching problems as well.

Person Re-ID approaches are based on either images [2, 1] or videos [26, 35]. Early works in person Re-ID were conducted in images, either via discriminative feature extraction [2, 34] or metric learning [1, 38, 17] approaches. Recently, various similar ideas have been proposed utilizing a deep learning setting [11, 12, 55, 17, 60]. However, Image-based approaches are intrinsically limited due to the visual ambiguity in inter-class appearance as well as the lack of spatio-temporal data. In contrast, Video-based Re-ID benefits from rich spatio-temporal data in video frames and addresses the task of matching between video sequences [35, 26]. This, combined with the release of large-scale datasets such as MARS [59] and DukeMTMC-VideoReID [54] has led to a gradual shift in the research community towards Video-based Re-ID from Image-based Re-ID.

Figure 1. Illustration of various solutions to focus on the subject rather than the background. (a) use of pose estimation [46, 58, 48] & (b) segmentation masks [39, 44] in Re-ID may miss salient accessories associated with the subjects (e.g., backpack, bag), (c) co-segmentation based attention (ours) exploits spatio-temporal data to capture common regions including persons along with their accessories (for e.g., a mobile phone).

Many of the video-based person Re-ID approaches in the literature extract frame-level features by considering the frame’s whole spatial area followed by temporal feature aggregation, i.e., LSTM/pooling [35, 62, 29]. One of the pioneering works by Melaughlin et al. [35] used a three-layer deep CNN to extract features from RGB & optical flow and a recurrent layer followed by a temporal average pooling (TP_{avg}) for feature aggregation. Chung et al. [8] extended this approach by utilizing a two-stream network. Unfortunately, such approaches often fail, especially in large-scale surveillance scenarios due to severe occlusions and background clutter. In such cases, it is highly probable that noisy background features from irrelevant non-salient regions may get misinterpreted as the person’s features and get aggregated in the video descriptor. Along with this, the subject alignment and scene variation aggravate the prob-
lem and result in a drastic drop in Re-ID accuracy.

Some works exploited augmented information such as pose & segmentation techniques to focus on the subject and avoid features from the background for generating an effective representation of the subject. One of the human pose estimation based approaches viz. Su et al. [46] proposed to use a Fully Convolutional Network (FCN)[31] based pose estimation model to extract part-based features. Similarly, Suh et al. [48] used OpenPose[3] model to accumulate part-based features for Re-ID. However, such methods have some drawbacks: (a) Albeit the pose estimation can effectively locate the person’s key joint locations (e.g., head, torso and legs), it misses out the salient accessories associated with the subject (e.g., backpack, bag, hat and coat) that are also important cues for Re-ID (Fig. 1(a)). (b) Standard pose estimation datasets may not cover the drastic viewpoint variations in surveillance scenarios, e.g., top-view (c) Surveillance images may not have sufficient resolution for stable pose estimation. Segmentation based Re-ID approaches were based on pre-trained models[40, 15]. For instance, Qi et al. [39] & Song et al. [44] explored the use of FCN[31] based pre-trained segmentation models to segment the subject. Again, these models are trained on datasets with segmentation masks only on humans and thus may not extract all parts of the subject including accessories (Fig. 1(b)).

Instead of using such expensive augmented information, an alternative solution is to use an Attention-driven approach, wherein the network is trained end-to-end. Li et al. [26] discovered a set of distinctive body parts using diverse spatial attentions and discriminative frames by a temporal attention model. Similarly, Wang et al. [52] computed features from automatically selected discriminative video fragments while simultaneously learning a video ranking function for person Re-ID. Although learned without explicit supervision, many of the attention-based approaches [52, 26, 52] are still sub-optimal since they work on “per-frame” basis, thus under-utilizing the rich spatio-temporal information available in video. Another recent line of approach [57, 5] tried to address this via Co-attention by leveraging inter-video (probe vs. gallery video snippets) co-attention. However, such approaches are computationally expensive and time consuming as such a processing has to be done for each probe-gallery instance pair separately.

In this paper, we propose a novel “Co-segmentation based Attention network” to effectively tackle the aforementioned problems in video-based Re-ID. Instead of a naïve “per-frame” or computationally intensive “inter-video” attention, we present an efficient intra-video attention inspired by Co-segmentation [50, 27] to jointly exploit the correlated information in multiple frames of the same video. As opposed to many of the existing heavily human-centric approaches (e.g., pose, segmentation), our approach relaxes the constraint by extracting task-relevant regions in the image that typically correspond to persons along with their accessories (Fig. 1(c)). To achieve co-segmentation, we propose a novel module named “Co-segmentation Activation Module” (COSAM) that effectively captures the attention between frames of a video. To the best of our knowledge, this work marks the first application of co-segmentation to Re-ID. Additionally, we conjecture that our intra-video attention mechanism may be useful in other video analytics applications also such as object tracking/segmentation and activity recognition. The primary contributions of this paper are:

- We propose a novel Co-segmentation inspired Re-ID architecture for video-based Re-ID.
- We formulate a plug-and-play “Co-segmentation Activation Module (COSAM)” that can be included in any deep neural architecture to enhance common abstract features and to suppress background features by jointly finding common features across frames.
- We visualize the co-segmentation based attention masks depicting the relevant frame regions, thus making our approach interpretable.

1.1. Related work on Object Co-segmentation

Based on the type of algorithm, the co-segmentation approaches are grouped into two categories: 1) Graph-based [4, 21, 25] and 2) Clustering-based [22, 49]. The former leveraged the shared structural representation among object instances from different images to jointly segment the common objects, whereas the latter motivates co-segmentation as a clustering task by grouping pixels/super-pixels in the common object regions. Classical approaches[43, 50] used hand-crafted features, such as SIFT [32] and HOG[9] for object instance representation, whereas the recent state-of-the-art methods are increasingly using deep learning approaches.

Recently, Li et al. [27] proposed a deep network to co-segment the regions by comparing their semantic similarity. Hsu et al. [18] proposed an unsupervised approach for co-segmenting the objects of a specific category without additional data annotations. Further, Chen et al. [6] presented an attention-based approach in the bottleneck layer of a deep neural network to activate semantically related features. Though having rich literature, the application of co-segmentation in other Computer Vision tasks is limited, and our work marks one of the first approaches endorsing the applicability of co-segmentation in other vision tasks.

2. A video-based Re-ID pipeline

In this paper, we follow a recent line of research yielding the current state-of-the-art in video-based Re-ID that can be summarized into a template framework as shown in Fig. 2. It consists of two primary components: (a) A Feature ex-
traction network: This is capable of extracting a meaningful abstract spatial representation from video frames either by hand-crafted features (SIFT, LBP, HoG, etc.) or automatically extracted deep CNN features by using pre-trained ImageNet models\cite{16, 19} such as ResNet and SE-ResNet. (b) Temporal feature aggregation: Here, the extracted frame-level feature vectors are aggregated to form a video-level feature vector to represent the person identity in the video. The complexity of feature aggregation techniques in the literature varies from a simple temporal pooling (TP\textsubscript{max/avg}) operation (average/max pooling) to complex temporal attention (TA) and recurrent layer (RNN) based aggregation \cite{13}. The aggregated video-level feature vectors are then used to compare (using $L_2$ distance or a learned metric) against other video instances for matching and retrieval purpose. In Table 1, we give a summary of prior work in video-based Re-ID using the aforementioned framework.

<table>
<thead>
<tr>
<th>Literature work</th>
<th>Feature extractor</th>
<th>Feature aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCN for Re-ID\cite{35}</td>
<td>Custom 3-layer CNN</td>
<td>RNN + TP\textsubscript{avg}</td>
</tr>
<tr>
<td>Two Stream Siamese\cite{1}</td>
<td>Custom 3-layer CNN</td>
<td>RNN + TP\textsubscript{avg}</td>
</tr>
<tr>
<td>Jointly attentive ST pooling\cite{50}</td>
<td>deep CNN + spatial pyramid pooling</td>
<td>Attentional TP</td>
</tr>
<tr>
<td>Comp. Snippet Sim.\cite{3}</td>
<td>ResNet-50</td>
<td>LSTM, Co-attentive embedding</td>
</tr>
<tr>
<td>Part-Aligned\cite{8}</td>
<td>GoogLeNet</td>
<td>Bilinear pooling</td>
</tr>
</tbody>
</table>

Table 1. A collection of approaches in the literature that are following the video-based Re-ID pipeline shown in Fig. 2.

A recent study by Gao \textit{et al.} \cite{13} re-visited the effect of various temporal aggregation layers with ResNet50 \cite{16} as the feature extractor. We extend this study by including yet another state-of-the-art architecture SE-ResNet50 \cite{19} and present the quantitative results in Table 2.

Based on our experiments in Table 2, we postulate certain key observations: \textbf{First}, the selection of the backbone network can influence the holistic system performance. This is quite noteworthy since not much research on the influence of the backbone network on Re-ID performance has been conducted. \textbf{Second}, it is observed that even a simple

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline
Feature extractor & Temp. Agg. & mAP & R1 & R5 & R20 & mAP & R1 & R5 & R20 \\
\hline
ResNet50\cite{13} & TP\textsubscript{avg} & 75.8 & 83.1 & 92.8 & 96.8 & 92.9 & 93.6 & \textbf{99.0} & \textbf{99.7} \\
ResNet50\cite{13} & TA & 76.7 & 83.3 & 93.8 & \textbf{97.4} & 93.2 & 93.9 & 98.9 & 99.5 \\
ResNet50\cite{13} & RNN & 73.8 & 81.6 & 92.8 & 96.7 & 88.1 & 88.7 & 97.6 & 99.3 \\
SE-ResNet50 & TP\textsubscript{avg} & 78.1 & 84.0 & 95.2 & 97.1 & 93.5 & 93.7 & \textbf{99.0} & \textbf{99.7} \\
SE-ResNet50 & TA & 77.7 & 84.2 & 94.7 & \textbf{97.4} & 93.1 & 94.2 & \textbf{99.0} & \textbf{99.7} \\
SE-ResNet50 & RNN & 75.7 & 83.1 & 93.6 & 96.0 & 92.4 & 94.0 & 98.4 & 99.1 \\
\hline
\end{tabular}
\caption{Evaluation of a simple video-based Re-ID framework on different feature extractor networks, feature aggregation techniques and datasets. Best results are shown in \textbf{Bold}.}
\end{table}

3. Co-segmentation for video-based Re-ID

Object co-segmentation is the task of identifying and segmenting common objects from two or more images according to “some” common characteristics \cite{50, 27} such as similarity of object-class and appearance. An illustration of co-segmentation is depicted in Fig. 3.

![Object Co-segmentation](image3.png)

Figure 3. An example illustration of object co-segmentation using images from the Caltech-UCSD Birds 200\cite{53} dataset.

As also noted in Section 1, a primary notion is to incorporate some common saliency associated with a person (along with his accessories) among video frames that can enhance the features from the person and suppress irrelevant background features. With this motivation, we exploit the co-segmentation inspired attention mechanism into the video-based person Re-ID task. The application of co-segmentation seems naturally relevant in video Re-ID since

\footnotesize{\textsuperscript{1}Winner of ILSVRC 2017 Image Classification Challenge\cite{10}
\textsuperscript{2}We investigate only the effect of temporal average pooling (TP\textsubscript{avg}) instead of max pooling (TP\textsubscript{max}), as the former is shown to be superior in \cite{35, 8, 13} }
the frames corresponding to a particular identity is known to contain a specific common object (person) of primary interest that is to be matched. In this regard, we propose a novel Co-Segmentation Activation Module (COSAM) layer (Section 4) that can be plugged between consecutive convolution blocks of a deep neural network.

Prior to explaining COSAM, we briefly review two deep network-based co-segmentation approaches that inspired our work. Li et al. [27] proposed an encoder-decoder Siamese architecture to co-segment the common objects by considering mutual correlations of spatial feature descriptors in the bottleneck layer of the encoder. By mutually correlating the feature descriptors between images at every spatial location, a correlation based cost matrix is computed and further passed to the decoder to estimate the co-segmentation mask. The same work also mentioned the idea of group co-segmentation to handle a group of images, simultaneously. Yet another work by Chen et al. [6] explored an approach in a Siamese encoder-decoder architecture to co-segment images based on common channel activations in the bottleneck layer. In particular, the notion of co-segmentation was achieved via conditioning the channel activations of one image on the channel activations of the other image (in image-pairs) and by taking average channel activation (in a group of images). Our COSAM layer is built upon the group co-segmentation approach from both of these papers, but reformulated for video Re-ID.

4. Co-segmentation activation module (COSAM)

We propose a Co-segmentation Activation Module (COSAM) that can be plugged between convolution blocks of several deep neural network architectures to induce the notion of co-segmentation. The architecture of the COSAM module is shown in Fig. 4. The input for the COSAM module is the set of frame-level feature maps of a person after a convolution block. The feature map is denoted by $F_{n,p} = CNN_L(I_{R}^{p})$, where $CNN_L$ refers to the network up-to the $L^{th}$ convolution block, $n$ is the index of the video frame ($1 \leq n \leq N$) of the person identified by the index $p$ & the feature maps are of dimension $D_L \times H_L \times W_L$ for each frame ($D_L = \text{Number of channels}$, $H_L = \text{Height}$, $W_L = \text{Width}$). Once the feature map enters COSAM, it undergoes a two-step process: (a) COSAM spatial attention (Section 4.1) & 2) COSAM channel attention (Section 4.2), that we detail next.

4.1. COSAM spatial attention

First, the input feature maps $\{F_{n,p}\}_{n=1}^{N}$ are passed through a dimension reduction layer $(1 \times 1$ convolution + BatchNorm[20] + ReLU[36]) to reduce the number of channels from $D_L$ to $D_R$ $(D_R << D_L)$. Thus, we get the feature maps of dimension $D_R \times H_L \times W_L$ as the output. The dimension reduction step is specifically carried out to speed-up the computations.

Our goal in the spatial attention step (Fig. 4 (a)) is to estimate a spatial mask for each frame belonging to a person that only activates the spatial locations of the person by consulting with all the given $N$ frames. In this regard, we build upon [27] such that given the spatial feature map $F_{n,p}$ of frame $I_{R}^{p}$ with dimension $D_R \times H_L \times W_L$, we consider the channel-wise feature vector at every spatial location $(i,j)$ $(1 \leq i \leq H_L, 1 \leq j \leq W_L)$ as a $D_R$ dimensional local descriptor of the frame at location $(i,j)$, denoted by $F_{n,i,j}^{(i,j)}$. To match the local regions across frames, for each frame $I_{R}^{p}$ and its location $(i,j)$, we compare the local descriptor $F_{n,i,j}^{(i,j)}$ to all the local descriptors of other $(N-1)$ frames available exhaustively. Here, the comparison is carried out using Normalized Cross Correlation (NCC) between the local de-
scriptors as it is robust to illumination variations and this was found to be more robust than a simple correlation[47]. The comparison results are reshaped into a 3D cost volume where each spatial location \((i, j)\) holds the comparison values. The idea of creating a cost volume in terms of matching the descriptors in an end-to-end learning framework has also been employed in other Computer Vision tasks such as geometric matching[42], image-based Re-ID[47] and stereo matching[23] among others.

Mathematically, it can be defined as:

\[
\text{Cost Volume}_{ij}(i, j) = \{ \text{NCC} \left( F_{n,p}^{(i,j)}, F_{m,p}^{(h,w)} \right) \mid \\
(1 \leq m \leq N, m \neq n) \\
(1 \leq h \leq H_L) \\
(1 \leq w \leq W_L) \}
\]  

(1)

Given two descriptors \(P, Q\) of \(D_R\) dimension, the NCC operation is defined as:

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

(2)

Here, \((\mu_P, \mu_Q)\) denote the mean of the descriptors \((P, Q)\), & \((\sigma_P, \sigma_Q)\) denote the standard deviations of the descriptors \((P, Q)\) respectively. (A small value of \(\epsilon = 1e^{-4}\) is added to \(\sigma\)’s to avoid numerical instability.

The cost volume is summarized by using a \(1 \times 1\) convolution layer followed by a sigmoid activation resulting in a spatial mask for the corresponding frame. The spatial mask is multiplied with the corresponding frame’s original input features \(F_{i,p}\) to activate only the local regions of images that are in consensus with all the \(N - 1\) frames. The output features after the spatial attention step are passed on to the channel attention step.

4.2. COSAM channel attention

In the channel attention step (Fig. 4 (b)), we intend to give more importance to the common important channels between the frames. To achieve this, we build upon [6] such that Global Average Pooling (GAP) is applied on the feature maps from the spatial attention step and the resulting feature vector is passed through a Multi-Layer Perceptron (MLP) followed by sigmoid activation to get the channel importance of each frame. The obtained channel importance vectors of all the \(N\) frames are average pooled together on each dimension to estimate the global channel-importance. The averaged channel importance vector is then multiplied with spatially-attended features to obtain the importance-weighted channel activations that are passed to the next layer.

5. Overall network architecture

Modern state-of-the-art image recognition network architectures (ResNet50, SE-ResNet, etc.) that are used as feature extractors in video-based Re-ID contain multiple consecutive \(CNN\) \(blocks\), in which the convolution layers are grouped according to the resolution of output feature maps: ResNet50 and SE-ResNet50 has five blocks (one initial convolution block followed by four consecutive Residual (or) Squeeze and Excitation (SE) residual blocks). We propose to plug in the COSAM layer after the \(CNN\) \(blocks\) in these network architectures. An illustration of proposed network architecture along with the COSAM layer is shown in Fig. 5. After getting the output of every \(CNN\) \(block\), the feature extractor employs a COSAM layer to co-segment the features and then the co-segmented features are passed to the next \(CNN\) \(block\). At the end of the feature extractor, the temporal aggregation layer (\(TP_{avg}\) or \(TA\) or \(RNN\)) is applied to summarize the frame-level descriptors to a video-level descriptor. The resulting video-level descriptor is used to predict the probability that the video belongs to a particular person identity.

5.1. Objective functions

For a fair comparison with the baseline[13] and due to their suitability for our task, we use the same loss functions as in [13]. The overall loss function can be written as:

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

(3)

Here, \(L_{CE}\) & \(L_{triplet}\) refer to the cross-entropy loss and batch triplet loss respectively and \(\lambda\) refers to the trade-off parameter between the losses (we use \(\lambda = 1\), as per [13]).
$B$ = batch size & $(I_i, I_{i+1}, I_{i-1})$ refer to the $i$th image in the batch and its hard positive and hard negative pair within the current batch, respectively.

**Cross-Entropy loss ($L_{CE}$):** This supervised loss is used to calculate the classification error among the identities. The number of nodes in the softmax layer depends on the number of identities in the training set.

**Batch triplet loss ($L_{triplet}$):** To reduce the intra-class variation and to increase the inter-class variation, the training instances are formed as a triplet where each triplet contains an anchor, a positive instance that belongs to the same class as the anchor and a negative instance that belongs to a different class than the anchor. Hard negative mining is carried out on the fly in each batch to select the hardest examples that pose a challenge for the model. Let $\{f_{I_A}, f_{I_+}, f_{I_-}\}$ be the video-level descriptors of a triplet, where $I_A, I_+, I_-$ are the anchor, positive and negative examples respectively. The triplet loss function is defined as:

$$L_{triplet}(I_A, I_+, I_-) = \max\{D(f_{I_A}, f_{I_+}) - D(f_{I_A}, f_{I_-}) + m, 0\} \quad (4)$$

Here, $m$ is the margin between the distances, $D(i, j)$ denotes the distance function between two descriptors $i, j$.

The Cross-entropy loss function is applied on the softmax probabilities obtained for the identities and the batch triplet loss is applied on the video-level descriptors to back-propagate the gradients.

### 6. Experiments

In this section, we evaluate the performance of the proposed COSAM layer by plugging it to two state-of-the-art deep architectures: ResNet50[16] & SE-ResNet50[19].

#### 6.1. Datasets and Evaluation protocol

We evaluate the proposed algorithm on three commonly used video-based person Re-ID datasets: MARS [59], DukeMTMC-VideoReID [54] and iLIDS-VID[51]. The MARS dataset[59] is the largest sequence-based person Re-ID dataset with 1261 identities and 20,478 video sequences, with multiple frames per person captured across 6 non-overlapping camera views. Among the total identities, 625 identities are used for training and the rest are used for testing. Additionally, 3,248 identities (disjoint with the train and test set) are used as distractors. DukeMTMC-VideoReID [54] is a subset of the DukeMTMC multicamera dataset [41], which was collected on outdoor scenario with varying viewpoint, illuminations, backgrounds and occlusions using 8 synchronized cameras. It contains 702 identities, each for training & testing, and 408 identities as the distractors. There are 369,656 tracklets for training, and 445,764 frames for testing & distractors. iLIDSLID [51] is a small dataset containing 600 sequences of 300 persons from two non-overlapping camera views. The sequences vary in length between 23 and 192 frames. As per the protocol followed in [51, 26], 10 random probe-gallery splits are used to perform experiments.

We use the standard evaluation metrics as followed in the literature[59, 26, 29, 48] viz., 1) **Cumulative Matching Characteristics (CMC)** & 2) **Mean average precision (mAP).** CMC is based on the retrieval capability of the algorithm to find the correct identity within the top-k ranked matches. CMC is used when only one gallery instance exists for every identity. We report rank-1, rank-5 and rank-20 CMC accuracies. The mAP metric is used to evaluate algorithms in multi-shot re-identification settings where multiple instances of same identities are present in the gallery.

#### 6.2. Implementation details

The proposed method is implemented using the PyTorch framework[37] and is available online3. During training, every video consists of $N = 4$ frames (as in baseline [13]) and each frame is of height = 256 and width = 128. The images are normalized using the RGB mean and standard deviation of ImageNet[10] before passing to the network. The network is trained using Adam optimizer with the following hyper-parameters : $\beta_1 = 0.9, \beta_2 = 0.999$, batch size $= 32$, initial learning rate $= 0.0001$, trade-off parameter between losses $\lambda = 1$ and COSAM dimension reduction size $D_R = 256$. We train the network for ~ 60K iterations and the learning rate is multiplied by 0.1 after every 15K iterations. The implementation was done in a machine with NVIDIA GeForce GTX 1080 Ti GPUs and takes around 8 hours to train a model with one GPU.

#### 6.3. Results & Discussion

In our experiments, every video of the person is split into multiple non-overlapping video-snippets of length $N$ frames and each snippet is passed through the network to obtain a snippet-level descriptor. Further, the video-snippet level descriptors are averaged to get the video-level descriptor. Then, these video-level descriptors are compared using the $L_2$ distance to calculate the CMC and mAP performances.

**Location of the COSAM layer within the network:** Without loss of generality, as a first step, the effect of the COSAM layer is evaluated by plugging it after each CNN block of the feature extractors and TP$_{avg}$ is used as the feature aggregation layer. The network is trained and evaluated on the MARS & DukeMTMC-VideoReID datasets and the quantitative results are shown in Table 3. From the results, it can be inferred that the inclusion of the COSAM module improves the baseline network and it is effective in the deeper layers (COSAM$_3$, COSAM$_4$, COSAM$_5$), as the features in those layers are more discriminative and abstract.

---

3 https://github.com/InnovArul/vidreid_cosegmentation

567
than the features at shallow layer(s). We also experiment with the inclusion of multiple COSAM blocks simultaneously. It is found that COSAM\textsubscript{4,5} (plugging in COSAM\textsubscript{4} & COSAM\textsubscript{5}) as in Fig. 5 achieves the best performance and is treated as our default proposed architecture in the rest of the experiments. An in-depth analysis by plugging in multiple COSAMs at various locations is detailed in the Supplementary Material.

<table>
<thead>
<tr>
<th>ResNet-50</th>
<th>mAP</th>
<th>R1</th>
<th>R5</th>
<th>R20</th>
<th>mAP</th>
<th>R1</th>
<th>R5</th>
<th>R20</th>
</tr>
</thead>
<tbody>
<tr>
<td>No COSAM [13]</td>
<td>75.8</td>
<td>83.1</td>
<td>92.8</td>
<td>96.8</td>
<td>92.9</td>
<td>93.6</td>
<td>99.0</td>
<td>99.7</td>
</tr>
<tr>
<td>COSAM\textsubscript{4}</td>
<td>68.3</td>
<td>77.7</td>
<td>90.1</td>
<td>96.1</td>
<td>88.9</td>
<td>90.2</td>
<td>98.4</td>
<td>99.0</td>
</tr>
<tr>
<td>COSAM\textsubscript{5}</td>
<td>76.9</td>
<td>82.7</td>
<td>94.3</td>
<td>97.3</td>
<td>93.6</td>
<td>94.0</td>
<td>98.7</td>
<td>99.9</td>
</tr>
<tr>
<td>COSAM\textsubscript{4,5}</td>
<td>76.8</td>
<td>82.9</td>
<td>94.2</td>
<td>97.1</td>
<td>93.8</td>
<td>94.7</td>
<td>98.7</td>
<td>99.9</td>
</tr>
<tr>
<td>COSAM\textsubscript{4,5}</td>
<td>76.6</td>
<td>82.8</td>
<td>93.9</td>
<td>97.2</td>
<td>93.2</td>
<td>93.7</td>
<td>98.4</td>
<td>99.9</td>
</tr>
</tbody>
</table>

Table 3. Evaluation of the backbone feature extractors with COSAM and temporal aggregation layer as TP\textsubscript{avg}. COSAM\textsubscript{i} implies plugging in COSAM layer after \textit{i}th CNN block.

**Visualizations:** To demonstrate the interpretability of our proposed method, we visualize the spatial attention masks of the COSAM\textsubscript{i} layer in SE-ResNet50+COSAM\textsubscript{4,5} model trained on MARS dataset (Fig. 6). The frames exhibit varying conditions such as scale, pose, viewpoint changes and partial occlusions. In Fig. 6(a), the predicted attention mask is able to focus on the person and avoid background features. In Fig. 6(b), despite the person occupying comparatively a small region of the frame, the COSAM layer still successfully focuses on the person based on task-relevant consensus. Although the buildings and trees are common in all the frames, our Co-segmentation inspired architecture specifically trained for Re-ID ignores the background regions. In Fig. 6(c), it can be observed that the spatial attention identifies the accessory (umbrella) carried by the person. Identifying the person with the aid of their belongings is one of the significant ways to discriminate the person by appearance. In Fig. 6(d), the partial occlusion scenario is handled successfully by avoiding the occluding object (cycle). More spatial mask illustrations are shown in Supplementary material.

**Effect of COSAM in the baseline model:** To understand the significance of the COSAM layer, we incorporate our best performing Co-segmentation based Re-ID module (COSAM\textsubscript{4,5}) into baseline video-based Re-ID pipelines with two feature extractors (ResNet50 and SE-ResNet50) and three different temporal aggregation layers (TP\textsubscript{avg}, TA, RNN) [13]. Table 4 represents the performance evaluation of the models. Our COSAM-based networks show consistent performance improvement (both CMC Rank and mAP) over the baseline models, in all three datasets. Between the backbone networks, SE-ResNet50 outperforms ResNet50 in both baselines and proposed case studies, highlighting the importance of a better backbone network selection. Among the temporal aggregation modules, although more or less similar performance is exhibited by TP\textsubscript{avg}, TA and RNN, the former (TP\textsubscript{avg}) results in the best mAP values in both MARS and DukeMTMC-VideoReID datasets & best CMC Rank-1 in iLIDS-VID. In particular, COSAM improves the mAP by 1.4% (ResNet50) &
6.4. Ablation studies

Effect of different frame lengths (N): We study the effect of the number of frames in a video on the performance of our best performing model. In particular, we analyze with frame lengths of $N = 2$, 4 and 8 in SE-ResNet50+COSAM$_{4,5}$+TP$_{avg}$ and the results are shown in Table 6. We found $N = 4$ frames to be optimal similar to [13]. Additionally, we also conduct studies comparing the effect of the frame selection scheme (Random vs. Sequence) and cross-dataset performance. We detail those experiments in the Supplementary material.

Attribute-wise performance gains: To understand the importance of COSAM in capturing attributes, we conduct attribute-wise empirical studies on the DukeMTMC-VideoReID dataset and present the results in Table 7. The significant improvements on attributes such as handbag, hat and backpack show that COSAM is indeed capturing the person’s attributes.

7. Conclusion and Future work

In this work, we proposed a novel “Co-segmentation inspired attention network” towards video-based Re-ID. In this regard, we presented a novel Co-segmentation based Attention Module (COSAM) for jointly learning the attention in the frames of a video to efficiently extract features in an end-to-end manner. In contrast to most existing Re-ID methods that exploit either pre-trained models and/or “per-frame” attention mechanism, the proposed model is able to extract the accessories also (e.g., bag, mobile phone, hat, umbrella) along with the persons, via task-relevant (Re-ID) attention across frames of the same video. Results show superior performance compared to the state-of-the-art. Such a co-segmentation based attention approach may be applied to other video-based Computer Vision problems also such as object tracking and video object segmentation.

Acknowledgements: This work is supported by grants from PM’s fellowship for Doctoral Research (SERB, India) & Google PhD Fellowship to Arulkumar Subramaniam.
References


