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.

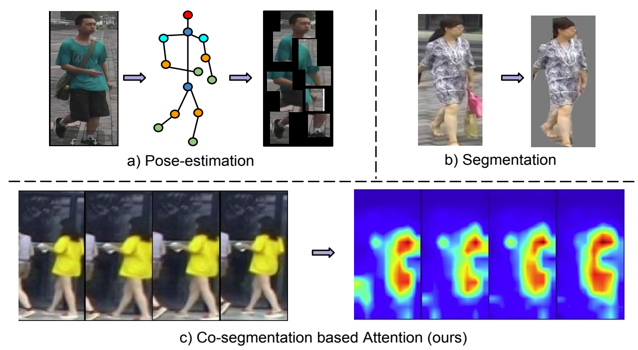


Figure 1. Illustration of various solutions to focus on the subject rather than the background. (a) use of pose estimation[46, 58, 48] & (b) segmentation masks[39, 44] in Re-ID may miss salient accessories associated with the subjects (*e.g.* backpack, bag), (c) co-segmentation based attention (ours) 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 Mclaughlin *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 problem 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 [62, 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-**

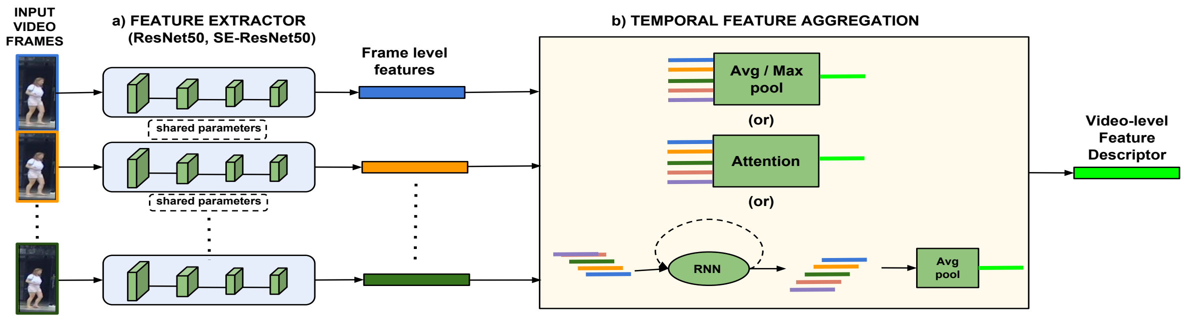


Figure 2. A standard Video-based Re-ID framework containing (a) Feature extractor and (b) Temporal feature aggregation components.

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[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$_{max/avg}$) operation (average/max pooling) to *complex* temporal attention (TA) and recurrent layer (RNN) based aggregation [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.

Literature work	Feature extractor	Feature aggregation
RCN for Re-ID[35]	Custom 3-layer CNN	RNN + TP$_{avg}$
Two Stream Siamese[8]	Custom 3-layer CNN	RNN + TP$_{avg}$
Jointly attentive ST pooling[56]	deep CNN + spatial pyramid pooling	Attentive TP
Comp. Snippet Sim.[5]	ResNet-50	LSTM, Co-attentive embedding
Part-Aligned[48]	GoogLeNet	Bilinear pooling

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 *et al*. [13] re-visited the effect of various temporal aggregation layers with ResNet50 [16] as the feature extractor. We extend this study by including yet another state-of-the-art architecture SE-ResNet50 [19][1] and present the quantitative results in Table 2.

Based on our experiments in Table 2, we postulate certain key observations: **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. **Second**, it is observed that even a simple

Feature extractor	Temp. Agg.	MARS				DukeMTMC-VideoReID			
		mAP	R1	R5	R20	mAP	R1	R5	R20
ResNet50[13]	TP$_{avg}$[2]	75.8	83.1	92.8	96.8	92.9	93.6	**99.0**	**99.7**
ResNet50[13]	TA	76.7	83.3	93.8	**97.4**	93.2	93.9	98.9	99.5
ResNet50[13]	RNN	73.8	81.6	92.8	96.7	88.1	88.7	97.6	99.3
SE-ResNet50	TP$_{avg}$	**78.1**	84.0	**95.2**	97.1	**93.5**	93.7	**99.0**	**99.7**
SE-ResNet50	TA	77.7	**84.2**	94.7	**97.4**	93.1	**94.2**	**99.0**	**99.7**
SE-ResNet50	RNN	75.7	83.1	93.6	96.0	92.4	94.0	98.4	99.1

Table 2. Evaluation of a simple video-based Re-ID framework on different feature extractor networks, feature aggregation techniques and datasets. Best results are shown in **Bold**.

TP$_{avg}$ layer performs on-par with complex attention/RNN based aggregation layers, as also reported in [7].

We incorporate our idea of Co-segmentation in this base architecture, and this is described next.

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 [50, 27] such as similarity of object-class and appearance. An illustration of co-segmentation is depicted in Fig. 3.

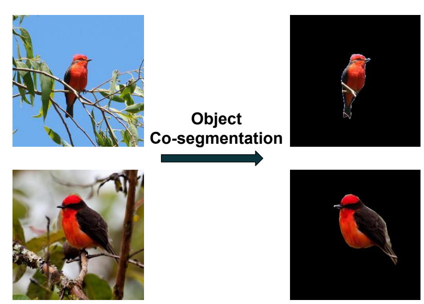


Figure 3. An example illustration of object co-segmentation using images from the Caltech-UCSD Birds 200[53] dataset.

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

[1] Winner of ILSVRC 2017 Image Classification Challenge[10]

[2] We investigate only the effect of temporal *average* pooling (TP$_{avg}$) instead of *max* pooling (TP$_{max}$), as the former is shown to be superior in [35, 8, 13]

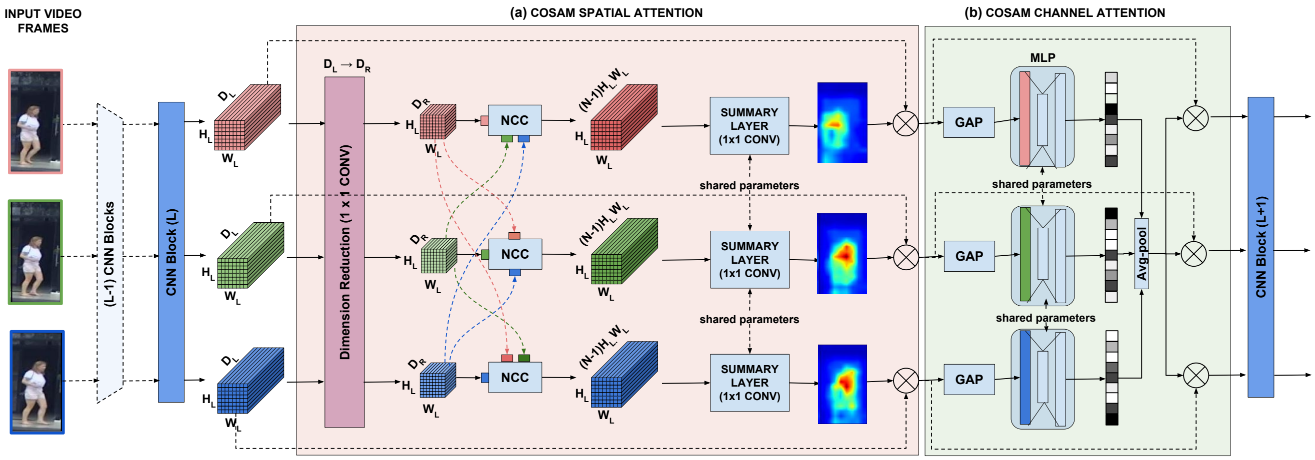


Figure 4. The Co-segmentation Activation Module (COSAM) plugged in-between L^{th} and $(L+1)^{th}$ CNN blocks (Best viewed in color). COSAM consists of two steps: (a) In the **COSAM spatial attention step**, after dimensionality reduction, the feature maps are passed through Normalized Cross Correlation (NCC) matching layer and a summarization layer to get the corresponding spatial attention mask. (b) In the **COSAM channel attention step**, the features are enriched by considering the strength of common channel activation.

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**-**s**egmentation **A**ctivation **M**odule (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 $F_{n,p} = CNN_L(I_n^p)$, 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 = Number of channels, H_L = Height, W_L = 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 (1×1 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_n^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 $F_{n,p}^{(i,j)}$. To match the local regions across frames, for each frame I_n^p and its location (i, j), we compare the local descriptor $F_{n,p}^{(i,j)}$ 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-

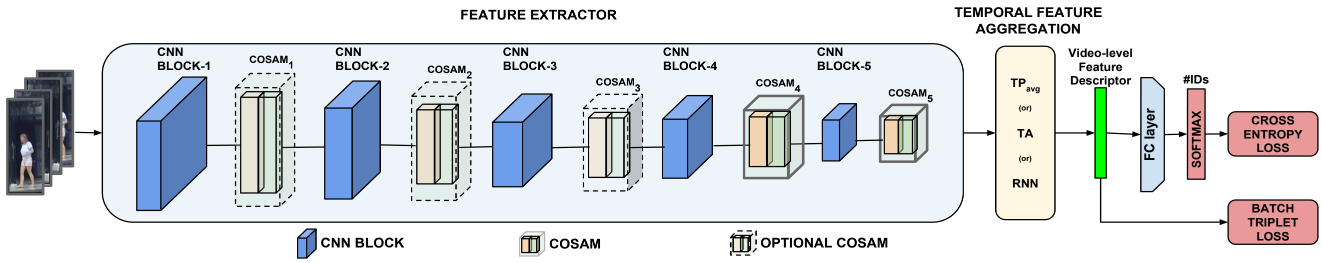


Figure 5. Illustration of overall architecture with the inclusion of the proposed COSAM layer. Here, two COSAM modules are plugged in after 4^{th} and 5^{th} CNN blocks; nevertheless, could be plugged in after any CNN block optionally (shown in dots). TP_{avg} = Temporal average pooling, *TA* = Temporal attention, *RNN* = Recurrent neural network and *FC layer* = Fully Connected layer.

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:

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

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

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

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. (A small value of $\epsilon = 1e^{-4}$ is added to $\sigma's$ to avoid numerical instability).

The cost volume is summarized by using a 1×1 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:

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

Here, L_{CE} & $L_{triplet}$ refer to the cross-entropy loss and batch triplet loss respectively and λ refers to the trade-off parameter between the losses (we use $\lambda = 1$, as per [13]),

B = batch size & (I_i, I_{i_+}, I_{i_-}) refer to the ith 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:

$$\begin{aligned} L_{triplet}(I_A, I_+, I_-) = \max\{D\left(f_{I_A}, f_{I_+}\right) \\ - D\left(f_{I_A}, f_{I_-}\right) + m, 0\} \end{aligned} \tag{4}$$

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 backpropagate 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 multi-camera dataset [41], which was collected on outdoor scenario with varying viewpoint, illuminations, background 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. iLIDS-LID [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

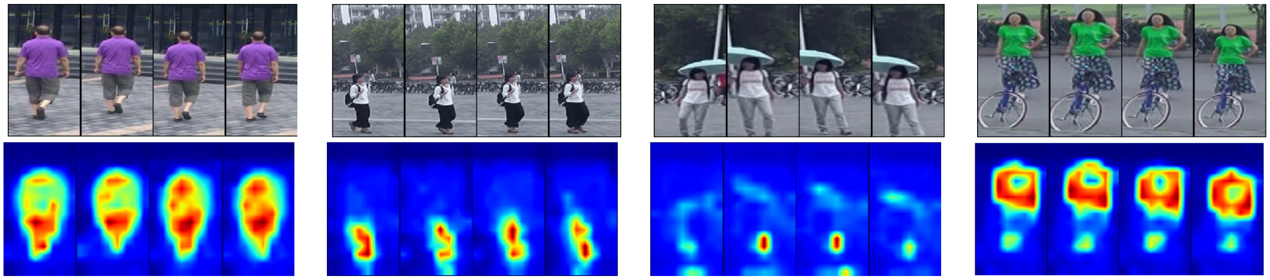
The proposed method is implemented using the PyTorch framework[37] and is available online[3]. During training, every video consists of $N = 4$ frames (as in baseline [13]) and each frame is of height = 256 and width = 128. The images are normalized using the RGB mean and standard deviation of ImageNet[10] before passing to the network. The network is trained using *Adam* optimizer with the following hyper-parameters : β_1 = 0.9, β_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 $\sim$ 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



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

Figure 6. Visualization of co-segmentations. The second row shows the segmentation maps corresponding to the images in the first row.

than the features at shallow layer(s). We also experiment with the inclusion of multiple COSAM blocks simultaneously. It is found that $\text{COSAM}_{4,5}$ (plugging in COSAM_4 & COSAM_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.

	COSAM_i	MARS				DukeMTMC-VideoReID			
		mAP	R1	R5	R20	mAP	R1	R5	R20
ResNet50	No COSAM [13]	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**
	$\text{COSAM}_{4,5}$	**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**
	$\text{COSAM}_{4,5}$	**79.9**	84.9	95.5	**97.9**	**94.1**	**95.4**	**99.3**	99.8

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

Visualizations: To demonstrate the interpretability of our proposed method, we visualize the spatial attention masks of the COSAM_4 layer in SE-ResNet50+$\text{COSAM}_{4,5}$ 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.

	Temp. Agg.	COSAM_i	MARS			Duke			iLIDS-VID	
			mAP	R1	R5	mAP	R1	R5	R1	R5
ResNet50	TP_{avg}[13]	-	75.8	83.1	92.8	92.9	93.6	99.0	73.9	92.6
	TP_{avg}	$\text{COSAM}_{4,5}$	**77.2**	**83.7**	94.1	**94.0**	94.4	99.1	**75.5**	94.1
	TA [13]	-	76.7	83.3	93.8	93.2	93.9	98.9	72.3	92.4
	TA	$\text{COSAM}_{4,5}$	76.9	83.6	93.7	93.4	**94.6**	98.9	74.9	94.4
	RNN[13]	-	73.8	81.6	92.8	88.1	88.7	97.6	68.5	93.2
	RNN	$\text{COSAM}_{4,5}$	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}	$\text{COSAM}_{4,5}$	**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	$\text{COSAM}_{4,5}$	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	$\text{COSAM}_{4,5}$	76.0	83.4	93.9	92.5	93.9	98.3	77.8	97.3

Table 4. Comparison of the baseline models with best performing COSAM-configuration ($\text{COSAM}_{4,5}$) along with different feature extractor networks, feature aggregation techniques and datasets. Here, $\text{COSAM}_{4,5}$ = COSAM layer is placed after 4^{th} and 5^{th} CNN blocks of the baseline model, Duke = DukeMTMC-VideoReID dataset. Best mAP & CMC Rank-1 per backbone network are shown in **red** and **blue** colors respectively. mAP is not applicable for iLIDS-VID due to single gallery instance per probe.

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 ($\text{COSAM}_{4,5}$) into baseline video-based Re-ID pipelines with two feature extractors (ResNet50 and SE-ResNet50) and three different temporal aggregation layers (TP_{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_{avg}, TA and RNN, the former (TP_{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) &

Network	Deep model?	MARS			
		mAP	R1	R5	R20
LOMO+XQDA[28]	No	16.4	30.7	46.6	60.9
JST-RNN[62]	Yes	50.7	70.6	90.0	97.6
QAN[30]	Yes	51.7	73.7	84.9	91.6
Context Aware Parts[24]	Yes	56.1	71.8	86.6	93.0
IDE+XQDA+ReRanking[61]	Yes	68.5	73.9	-	-
TriNet [17]	Yes	67.7	79.8	91.4	-
Region QEN[45]	Yes	71.1	77.8	88.8	94.1
Comp. Snippet Sim.[5]	Yes	69.4	81.2	92.1	-
Part-Aligned[48]	Yes	72.2	83.0	92.8	96.8
RevisitTempPool[13]	Yes	76.7	83.3	93.8	97.4
[13] + 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[61]	Yes	**87.4**	**86.9**	**95.5**	**98.0**

Network	Deep model?	DukeMTMC-VideoReID			
		mAP	R1	R5	R20
ETAP-Net[48]	Yes	78.34	83.62	94.59	97.58
RevisitTempPool[13]	Yes	93.2	93.9	98.9	99.5
[13] + 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 5. Comparison of our best model with state-of-the-art methods on MARS & DukeMTMC-VideoReID datasets.

1.8% (SE-ResNet50) in MARS and 1.1%(ResNet50) & 0.6% (SE-ResNet50) in DukeMTMC-VideoReID respectively. Regarding the CMC Rank, we observe an improvement of 0.6% (ResNet50) & 0.9% (SE-ResNet50) in MARS, 0.8% (ResNet50) & 1.7% (SE-ResNet50) in DukeMTMC-VideoReID and 1.6% (ResNet50) & 2.7% (SE-ResNet50) in iLIDS-VID.

Comparison with state-of-the-art methods We compare our method with the state-of-the-arts [28, 17, 30, 24, 61, 17, 45, 5, 48, 13] in MARS and DukeMTMC-VideoReID datasets and the results are shown in Table 5. It is observed that our proposed COSAM module applied in SE-ResNet50 (COSAM$_{4,5}$) along with TP$_{avg}$ achieves the best performance. In particular, our approach has $\sim$0.9% improvement in CMC Rank-1 as well as 1.8% improvement in mAP in the MARS dataset over the best performing method ([13] + SE-ResNet50 + TP$_{avg}$). Apart from this, applying re-ranking[61] further increases the performance to +2.0% CMC Rank-1 and +7.5% mAP. Intuitively, such an improved CMC Rank-1 (86.9%) shows that majority of the subjects are correctly identified in the first rank, whereas the improved mAP result (87.4%) denotes that multiple instances of the person are ranked precisely at the top in a multi-shot setting that is significant in retrieval problems. We attribute this improvement to the effectiveness of the COSAM layer in suppressing noise and aiding the network learn about identifying relevant common objects. Similarly, our COSAM layer with SE-ResNet50 achieves 0.6% improvement in mAP and 1.7% improvement in CMC Rank-1 with DukeMTMC-VideoReID dataset. The performance comparison of iLIDS-VID dataset is shown in the supplementary material.

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.

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 6. Evaluation of the influence of track length T on Re-ID performance of the best performing model *SE-ResNet50+COSAM$_{4,5}$+TP$_{avg}$*.

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.

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

Table 7. Attribute-wise perf. comparison on Duke reveals the effectiveness of COSAM to capture features of person's accessories. Here, R50=ResNet50, SE50=SE-ResNet50, C$_{4,5}$=COSAM$_{4,5}$.

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-s*egmentation based *A*ttention *M*odule (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

[1] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2990–2999, 2017. 1

[2] Loris Bazzani, Marco Cristani, and Vittorio Murino. Sdalf: modeling human appearance with symmetry-driven accumulation of local features. In *Person re-identification*, pages 43–69. Springer, 2014. 1

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 2

[4] Haw-Shiuan Chang and Yu-Chiang Frank Wang. Optimizing the decomposition for multiple foreground cosegmentation. *Computer Vision and Image Understanding*, 141:18–27, 2015. 2

[5] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018. 2, 3, 8

[6] Hong Chen, Yifei Huang, and Hideki Nakayama. Semantic aware attention based deep object co-segmentation. *arXiv preprint arXiv:1810.06859*, 2018. 2, 4, 5

[7] Qiming Chen and Ren Wu. Cnn is all you need. *arXiv preprint arXiv:1712.09662*, 2017. 3

[8] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1983–1991, 2017. 1, 3

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005. 2

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3, 6

[11] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 1

[12] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. *arXiv preprint arXiv:1804.05275*, 2018. 1

[13] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *CoRR*, abs/1805.02104, 2018. 3, 5, 6, 7, 8

[14] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person Re-Identification*. Springer Publishing Company, Incorporated, 2014. 1

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 8

[18] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *IJCAI*, pages 748–756, 2018. 2

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3, 6

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 4

[21] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Image co-segmentation via saliency co-fusion. *IEEE Transactions on Multimedia*, 18(9):1896–1909, 2016. 2

[22] Armand Joulin, Francis Bach, and Jean Ponce. Multi-class cosegmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 542–549. IEEE, 2012. 2

[23] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 5

[24] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017. 8

[25] Lina Li, Zhi Liu, and Jian Zhang. Unsupervised image co-segmentation via guidance of simple images. *Neurocomputing*, 275:1650–1661, 2018. 2

[26] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018. 1, 2, 6

[27] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. *arXiv preprint arXiv:1804.06423*, 2018. 2, 3, 4

[28] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 8

[29] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *IEEE transactions on circuits and systems for video technology*, 28(10):2788–2802, 2018. 1, 6

[30] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 5790–5799, 2017. 8

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[32] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[33] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE, 2009. 1

[34] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *European Conference on Computer Vision*, pages 413–422. Springer, 2012. 1

[35] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016. 1, 3

[36] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 4

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[38] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010. 1

[39] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864*, 2018. 1, 2

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[41] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 6

[42] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6148–6157, 2017. 5

[43] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 993–1000. IEEE, 2006. 2

[44] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 1, 2

[45] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 8

[46] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3960–3969, 2017. 1, 2

[47] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, pages 2667–2675, 2016. 5

[48] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018. 1, 2, 3, 6, 8

[49] Zhiqiang Tao, Hongfu Liu, Huazhu Fu, and Yun Fu. Image cosegmentation via saliency-guided constrained clustering with cosine similarity. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[50] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR 2011*, pages 2217–2224. IEEE, 2011. 2, 3

[51] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European conference on computer vision*, pages 688–703. Springer, 2014. 6

[52] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016. 2

[53] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 3

[54] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5177–5186, 2018. 1, 6

[55] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 1

[56] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan 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. 3

[57] Ruimao Zhang, Hongbin Sun, Jingyu Li, Yuying Ge, Liang Lin, Ping Luo, and Xiaogang Wang. Scan: Self-and-collaborative attention network for video person re-identification. *arXiv preprint arXiv:1807.05688*, 2018. 2

[58] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 1

[59] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 1, 6

[60] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018. 1

[61] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 8

[62] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu 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. 1, 2, 8