Image and Vision Computing 113 (2021) 104246

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Flow guided mutual attention for person re-identification

Madhu Kiran [a,*], Amran Bhuiyan [a], Le Thanh Nguyen-Meidine [a,*], Louis-Antoine Blais-Morin [b], Ismail Ben Ayed [a], Eric Granger [a]

[a] Laboratoire d'imagerie, de vision et d'intelligence artificielle, Dept. of Systems Engineering, École de technologie supérieure, Montreal, Canada
[b] Genetec Inc., Montreal, Canada

ARTICLE INFO

Article history:
Received 27 November 2020
Received in revised form 23 May 2021
Accepted 25 May 2021
Available online 7 July 2021

Keywords:
Video surveillance
Person re-identification
Optical flow
Metric learning
Attention mechanisms

ABSTRACT

Person Re-Identification (ReID) is a challenging problem in many video analytics and surveillance applications, where a person's identity must be associated across a distributed non-overlapping network of cameras. Video-based person ReID has recently gained much interest given the potential for capturing discriminant spatio-temporal information from video clips that is unavailable for image-based ReID. Despite recent advances, deep learning (DL) models for video ReID often fail to leverage this information to improve the robustness of feature representations. In this paper, the motion pattern of a person is explored as an additional cue for ReID. In particular, a flow-guided Mutual Attention network is proposed for fusion of bounding box and optical flow sequences over tracklets using any 2D-CNN backbone, allowing to encode temporal information along with spatial appearance information. Our Mutual Attention network relies on the joint spatial attention between image and optical flow feature maps to activate a common set of salient features. In addition to flow-guided attention, we introduce a method to aggregate features from longer input streams for better video sequence-level representation. Our extensive experiments on three challenging video ReID datasets indicate that using the proposed approach allows to improve recognition accuracy considerably with respect to conventional gated-attention networks, and state-of-the-art methods for video-based person ReID.

1. Introduction

Person Re-Identification (ReID) refers to the problem of associating individuals over a set of non-overlapping camera views. It is a key object recognition tasks, that has recently drawn a significant attention due to its wide range of monitoring and surveillance applications, e.g., multi-camera target tracking, pedestrian tracking in autonomous driving, access control in biometrics, search and retrieval in video surveillance, and human-computer interaction communities. Despite the recent progress with deep learning (DL) models, person re-identification remains a challenging task due to the non-rigid structure of the human body, the variability of capture conditions (e.g., pose, illumination, blur), occlusions, and background clutter.

ReID systems can apply in image-based and video-based settings. State-of-the-art [1–4,14,36,38,49,50] approaches on image-based setting seek to associate still images of individuals captured with a network of non-overlapping cameras. In case of video-based ReID, input video tracklets of an individual are matched against a gallery of tracklet representations, captured with different non-overlapping cameras. A tracklet corresponds to a sequence of bounding boxes that were captured over time for a same person in a camera viewpoint, and are obtained using a person detector and tracker. Compared to image data, video data provides motion information in addition to appearance information which can further enable the system to capture person's body silhouette via post processing. Thus, video-based approaches allow to exploit spatio-temporal information (appearance and motion) for discriminative feature representation.

As illustrated in Fig. 1, state-of-the-art approaches for video-based person ReID typically learn global features in an end-to-end fashion, through various temporal feature aggregation techniques [16,17,19,46]. From this figure, the query input to the feature extractor is a video clip (i.e, a set of consecutive bounding boxes extracted from a tracklet) of N frames long. A single feature vector is extracted by aggregating features from each frame in the query video clip. This is then compared with a gallery containing n identities, each one having a set of aggregated feature representations of clips from previously-captured tacklets.

Given a video clip (fixed size set of bounding boxes extracted from a tracklet), the feature extractor (CNN backbone) produces image-level features, while the feature aggregator generates a single feature representation at the clip level, using either average pooling, weighted addition, max pooling, recurrent NNs, etc. [16] in the temporal domain.

* Corresponding authors.
E-mail addresses: madhu_sajc@hotmail.com (M. Kiran), amran.apece@gmail.com (A. Bhuiyan), le-thanh.nguyen-meidine.1@ens.etsmtl.ca (L.T. Nguyen-Meidine), lablaismorin@genetec.com (L.-A. Blais-Morin), Ismail.BenAyed@etsmtl.ca (I.B. Ayed), Eric.Granger@etsmtl.ca (E. Granger).

https://doi.org/10.1016/j.imavis.2021.104246
0262-8856/

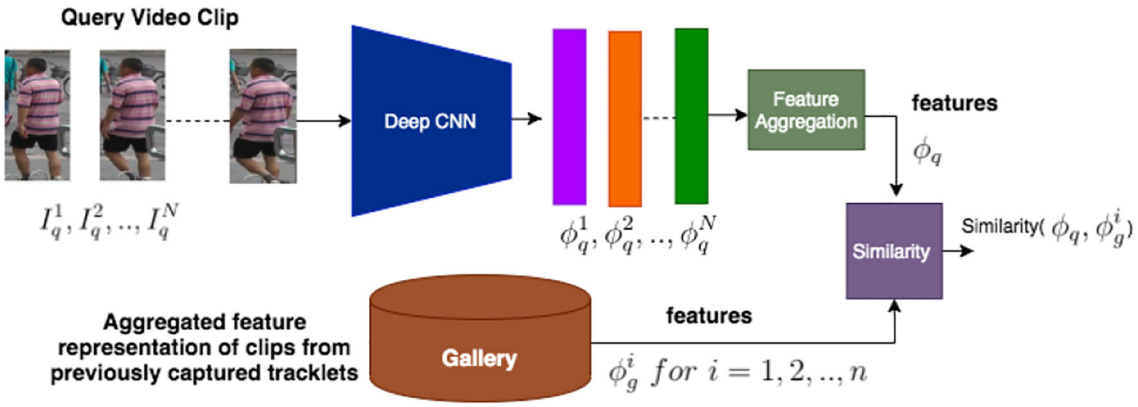

Fig. 1. Block diagram of a generic DL model specialized for video-based person ReID. Each query video clip from a non-overlapping camera is input to a backbone CNN to produce a set of features embeddings, one per bounding box image. The features are then aggregated to produce an aggregated feature representation each for both video and optical flow clip, which is then matched against clip representations stored in the gallery.

Although these aggregation approaches enable to incorporate diverse tracklet information for matching, and can achieve a higher level of accuracy than image-based approaches, they often fail to efficiently capture temporal information which could propagate as salient features throughout the video sequence. Additionally, the performance of state-of-the-art methods decline as the length of video clips grows beyond 4 or 6 frames [16,46].

Optical flow streams has been previously used to capture the motion dynamics of a person walking in a video stream. Moreover, as shown in Fig. 2, the visual appearance of optical flow of a person walking or moving resembles the silhouette of the person, often suppressing the background static objects. This can potentially be used as a mask on the appearance stream to highlight common saliency between frames. In addition to it highlighting common saliency across frames, the

Fig. 2. Example of a sequence of bounding boxes images from the MARS dataset (top row), and its corresponding dense optical flow map (bottom row). The common saliency in the sequence can be observed from the optical flow map.

silhouette produced by optical flow can therefore be a good source of spatio-temporal attention.

Fig. 2 shows that the flow features are coarse representation of semantic information of moving objects. Unlike action recognition which depends heavily on motion features, accurate ReID is more dependent on appearance features. Hence, there is a scope to combine the strengths of optical flow and appearance (video stream) features for ReID. Previous attempts to include optical flow information into ReID systems [12,34,51] focused on integrating this information just as an additional input stream without exploiting the relationship further between these two streams. This approach has limited effectiveness because optical flow only represents coarse semantic features of moving objects (different from the image stream), and not image-like appearance information. Moreover, the model in [12] is related to the two-stream network proposed in [42] that incorporates motion and appearance feature for action recognition. Two-stream networks that are effective for action recognition are less effective for ReID [12].

Given the aforementioned justification, we consider the correlation between the visual appearances across motion and appearance streams, along with their individual contribution to motion dynamics. In order to capture long-term spatial information and temporal dynamics in a video clip, a method to aggregate features from longer sequence effectively is presented. This is unexplored in the literature for video person-ReID, thereby undermining the global saliency in the feature representation by using optical flow for both appearance and motion information.

In this paper, a DL model for flow-guided attention is introduced for video-based person ReID that encodes joint spatial attention between features of temporal (optical flow) and spatial (image appearance) information across a tracklet. The proposed Mutual Attention network enables to jointly learn a feature embedding that incorporates relevant spatial information from human appearance, along with their motion information, from both appearance and motion streams. The Mutual Attention network includes both optical flow stream and image stream for ReID, and leverages the mutual appearance and motion information. We also propose a feature aggregation method that allows integrating information aggregation from longer tracklets. Unlike prior work in literature where feature aggregation is achieved by pooling or temporal attention from image feature, the proposed Mutual Attention network relies on a weighted feature addition method over images in a sequence to produce a single feature descriptor based on optical flow and image information. During feature aggregation, a reference frame from each tracklet is selected based on maximum activation from both streams, and weights are assigned for individual features using image and optical flow information. Attention is enabled from optical flow in both spatial and temporal domain to extract discriminant features for ReID.

The paper is organized as follows. The next section provides some background on DL models for spatio-temporal recognition, optical flow, and attention mechanisms, as they relate to video-based person ReID. Then, Section 3 describes the architecture and associated formulation for our Mutual Attention network. Section 4 and 5 present the experimental methodology and qualitative and quantitative results, respectively. The proposed flow-guided mutual attention network is validated on the challenging MARS, Duke-MTMC, and ILIDS-VID datasets for video-based person ReID. Experimental results show that it can outperform state-of-the-art approaches. Results also indicate the its potential for higher accuracy by processing longer video clips to capture multiple appearance variations.

2. Related work

2.1. Image-based person-ReID

The idea of using CNNs for ReID stems from Siamese Network [5], which involves two sub-networks with shared weights, and is suitable for finding the pair-wise similarity between query and reference images. It has first been used in [56] that employs three Siamese sub-networks for deep feature learning. Since then many authors focus on designing various DL architectures to learn discriminative feature embedding. Most of these deep-architecture based ReID [1,9,10,30,52] approaches introduce an end-to-end ReID framework, where both feature embedding and metric learning have been investigated as a joint learning problem. In [1,52], a new layer is proposed to capture the local relationship between two images, which helps modeling pose and viewpoint variations in cross-view pedestrian images. Recent ReID approaches [2,37,40,45,47,49,59,60,62] rely on incorporating contextual information into the base deep ReID model, where local and global feature representations are combined to improve accuracy. [23,35] use dissimilarity space instead of feature space for domain adaptation and occluded re-ID. A few attention-based approaches for deep re-ID [24, 45,60] address misalignment challenges by incorporating a regional attention sub-network into a base re-ID model. A thorough review of state-of-the-art on architecture-based approaches underscores the importance of considering local representations, e.g., by dividing the image into soft stripes [49] or by pose-based part representation [37, 40,45,47,59,60,62]. Although these methods have achieved considerable performance improvements, they fail to incorporate temporal information due to their image-based setting.

2.2. Video-based person-ReID

Video ReID has recently attracted some interest since temporal information allows dealing with ambiguities such as occlusion and background noise [16,17,19,46]. An important problem in video-based ReID is the task of aggregating the image level features to obtain one single composite feature or descriptor for a video sequence. [16] have approached this problem by frame level feature extraction and temporal fusion by using recurrent NNs (RNNs), average pooling, and temporal attention (based on image features). Average Pooling in temporal dimension can be viewed as summing the features of the sequence by giving equal and normalized weights to them. Average pooling of image instance features from a given sequence have proved to be useful in most of the cases, even compared to other DL model based on RNNs or 3D-CNN [16]. 3DCNN has been experimented in [16,26] but have not been very effective in summarizing video sequence for reID. But there could be certain case of individual image in a sequence such that they either have higher noise content or the appearance in the image does not contribute much to an individual's identity, then these become the debatable cases for Average Pooling.

2.3. Attention mechanisms

Attention can be interpreted as a means of biasing the allocation of available computational resources towards the most informative components of a signal [20]. A mask guided attention mechanism has been proposed in [43], where a binary body mask is used in conjunction with the corresponding person image to reduce background clutter. Somewhat similar to [43], co-segmentation networks have achieved significant improvements in ReID accuracy over the baseline by connecting a new COSAM module between different layers of a deep feature extraction network [46]. Co-Segmentation allows extracting common saliency between images, and using this information for both spatial- and channel-wise attention. Other related work for attention in video ReID, [7] attention is employed in both temporal and and spatial domain. Video stream has been taken advantage of by [44] by extracting complementary region based feature by from different frames to obtain informative features as a whole.

2.4. Optical flow as temporal stream

It often serves as a good approximation of the true physical motion projected onto the image plane [51]. Optical flow has been employed for temporal information fusion in [12,34], in a two stream Siamese

Network with a weighted cost function to combine the information from both the streams. It uses a CNN that accepts both optical flow and color channels as input, and a recurrent layer to exploit temporal relations. Its important to note that prior to [12,42] have used two stream networks but for action recognition. Two stream networks on their own are useful in action recognition as impact of motion cues in action recognition are higher in action recognition than that of ReID [12]. Therefore here is a necessity to use optical flow in a way that it can be leveraged for appearance related task. However, traditional two-stream networks are unable to exploit a critical component in re-id i.e. appearance across both optical flow and image stream together. Similar to our motivation for using optical flow for appearance along with motion has been discussed in [32]. Similar to our work, motivation for considering long term temporal relationship has been discussed in [11].

It can be summarized from the above that, as discussed in [41,43,46], various saliency feature enhancement methods have been proposed, and they often improve the overall accuracy. Optical flow typically encodes motion information in contrast to appearance information, and hence there is scope to enhancing appearance information from motion and vice-versa. The advantages of long-term information fusion across tracklets are highlighted in [11,32]. Although long-term aggregation of tracklet information can be beneficial, it also suffers from integration of noise while processing longer tracklets. Our proposed approach aims to mitigate this problem through attention-based feature aggregation, thereby taking full advantage of long-term feature aggregation.

3. Proposed mutual attention network

We propose a new model for flow-guided attention – the Mutual Attention network. It learns spatial–temporal attention from optical flow thereby focusing on common salient features of a given person during its motion across consecutive frames of a given video clip. Although a two stream Siamese network has been proposed in [12], they have included optical flow as an input for re-identification, and do not exploit the full potential of this information. Hence, we seek to leverage the visual appearance of both spatial and temporal streams, i.e., image and optical flow streams, by producing a correlation map between them in the feature space. This correlation map provides attention for both input streams. The temporal information in both the streams is enhanced by enabling the use of longer video and optical flow clips with our proposed feature aggregation method. Our Mutual Attention network therefore includes both optical flow and image streams to produce mutual attention, and also to combine the features into an aggregated representation from a video or optical flow tracklet or clip.

As illustrated in Fig. 3, the network accepts two streams of input (optical flow and image sequences). At the last layer of the network, the features from the two streams are concatenated after feature aggregation in the temporal domain. While the image stream helps in ReID by focusing on the appearance of the person, optical flow stream helps by capturing motion pattern of a given person. We propose to achieve feature aggregation to produce a feature vector by weighted addition. Our proposed method handles generation of weights that indicate the importance of individual image feature in producing a single video feature leveraging upon mutual attention.

3.1. Mutual attention

In contrast to the previous method for flow guided attention, we propose to produce cross-stream attention or mutual attention between the optical flow stream and image stream to emphasize areas in the feature space that have high activation across both the streams.

An input video clip is represented by $\mathbf{I_c^1}$,$\mathbf{I_c^2}$, …,$\mathbf{I_c^n}$ and corresponding optical flow estimations $\mathbf{F_c^1}$,$\mathbf{F_c^2}$, …,$\mathbf{F_c^n}$, where c indicates the identity of the video clip of length n, our objective is to extract a discriminative feature vector $\boldsymbol{\phi_c}$ for ReID. Given a video clip and its corresponding flow maps, we extract the features $\boldsymbol{\phi_c}$ and $\mathbf{f_c}$ from the deep CNNs respectively. The expected output is a concatenated feature vector of both optical flow and image features to be used for ReID. Both the CNNs share common architecture but do not share the parameters. Let l be the intermediate layer of the k-layer deep CNN and let appearance CNN be represented by H_{app} and optical flow stream CNN by H_{flow} with a total number of k layers. With $t = 1,2,3\ldots n$ we have:

$$\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{t}} = H_{\text{app}}\left(\mathbf{I}_{\mathbf{c}}^{\mathbf{t}}\right) \;,\; \mathbf{f}_{\mathbf{c}}^{\mathbf{t}} = H_{\text{flow}}\left(\mathbf{F}_{\mathbf{c}}^{\mathbf{t}}\right) \tag{1}$$

If features from layer l are expressed as $\phi^{\mathbf{l}}$, then:

$$\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{l,t}} = H_{\text{app,l}}\left(\mathbf{I}_{\mathbf{c}}^{\mathbf{t}}\right) \;,\; \mathbf{f}_{\mathbf{c}}^{\mathbf{l,t}} = H_{\text{flow,l}}\left(\mathbf{F}_{\mathbf{c}}^{\mathbf{t}}\right) \tag{2}$$

Both the features at layer l are of dimensions, $N \times C \times I \times J$, representing sequence length,channels, width, height respectively. The features are then passed through 1x1 convolution with *Relu* activation to produce a map of size $N \times 1 \times I \times J$ each. The correlation between the features is given by:

$$\boldsymbol{\rho} = \zeta_{\text{app}}\left(\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{l,t}}\right) \odot \zeta_{\text{flow}}\left(\mathbf{f}_{\mathbf{c}}^{\mathbf{l,t}}\right) \tag{3}$$

In the Eq. (3), ζ_{app} and ζ_{flow} are the embeddings with 1x1 convolution with *Relu* discussed above. $\boldsymbol{\rho}$ when activated by a sigmoid function forms the mutual attention map M_c^t between both the streams of input is:

$$\mathbf{M}_{\mathbf{c}}^{\mathbf{t}} = \frac{1}{1 + e^{-\rho}} \tag{4}$$

Eq. (3) aims to capture spatial similarities between the Deep CNN features of optical flow and video stream. Optical flow is the result of motion across two frames hence optical flow activation contains information of common saliency between frames. It mostly represents a silhouette of the person. Optical flow could also contain some noisy background motion. Activation from the video stream contains Person specific spatial information along with noises such as occlusion. The noise in both these streams is not common, but the activations in salient regions are common. Multiplying these two activations as in the Eq. (3) identifies similar salient regions in both the streams, thereby reducing the activations due to noise. Therefore, the product fortifies the activations across common saliency in both the streams.

Finally, mutual attention is applied to the intermediate features $\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{l,t}}$ and $\mathbf{f}_{\mathbf{c}}^{\mathbf{l,t}}$ at the intermediate layer (by an element-wise multiplication of attention map with feature maps) to obtain mutually attended appearance features $\Psi_{\boldsymbol{app}}$ and $\Psi_{\boldsymbol{flow}}$ to continue feature extraction continues in the remaining layers of the deep CNN to obtain final output features $\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{t}}$ and $\mathbf{f}_{\mathbf{c}}^{\mathbf{t}}$ for image and flow stream, respectively:

$$\Psi_{\boldsymbol{app}}{}_c^{\,t} = \boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{l,t}} \odot \mathbf{M}_{\mathbf{c}}^{\mathbf{t}} \;,\; \Psi_{\boldsymbol{flow}}{}_c^{\,t} = \mathbf{f}_{\mathbf{c}}^{\mathbf{l,t}} \odot \mathbf{M}_{\mathbf{c}}^{\mathbf{t}} \tag{5}$$

3.2. Weighted feature addition

We propose a method to aggregate image level features to obtain a single feature vector for a given video sequence particularly enabling to use longer video sequences. The appearance features and optical flow features are then concatenated to for ReID during inference, and to learn a classifier during training.

The output from image and optical flow stream CNNs generate $\boldsymbol{\phi}_{\mathbf{c}}$ for a sequence c from instances $\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{1}}$,$\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{2}}$, …,$\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{n}}$ and $\mathbf{f_c}$ from $\mathbf{f}_{\mathbf{c}}^{\mathbf{1}}$,$\mathbf{f}_{\mathbf{c}}^{\mathbf{2}}$, …,$\mathbf{f}_{\mathbf{c}}^{\mathbf{n}}$. The first task is to identify salient feature from a given sequence of features. In our case, a salient feature can be defined as the one that has maximum activation in both image and flow stream. Since the features have been attended by mutual attention, given a sequence, a max operation in the temporal domain for each of the sequence will identify the salient feature among the sequence. We hereafter will refer to this salient feature

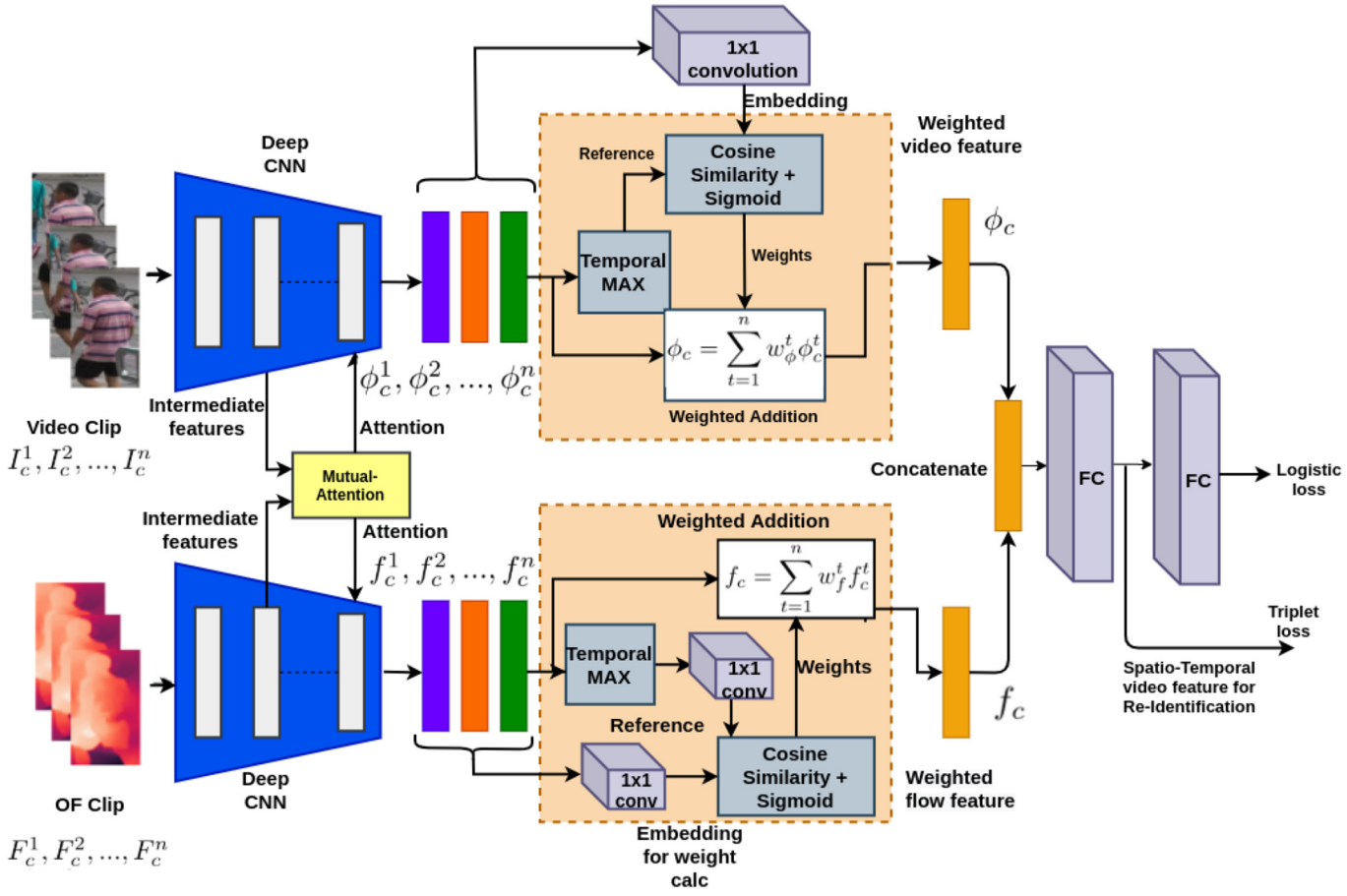

Fig. 3. Our proposed Mutual Attention network architecture. The system inputs is a video clip and corresponding flow maps. The corresponding features of the inputs attend to each other using our Mutual Attention module. The network outputs a concatenated feature representation from both optical flow and image stream.

as reference frame denoted by $\boldsymbol{\phi}_{\mathbf{c}}^{max}$ and $\mathbf{f}_{\mathbf{c}}^{\mathbf{max}}$. In the next step an adaptive weight is generated for each of the features in the sequence based on how close each feature is with the reference feature. This is achieved by applying a cosine similarity between the reference feature and rest of the features int he sequence. The cosine similarity function is not applied directly on the features $\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{n}}$ and $\mathbf{f}_{\mathbf{c}}^{\mathbf{n}}$. Instead a tiny embedding $\boldsymbol{\varepsilon}(.)$ is applied on the $\boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{n}}, f_c^n$ and reference feature $\boldsymbol{\phi}_{\mathbf{c}}^{max}$, $\mathbf{f}_{\mathbf{c}}^{\mathbf{max}}$ to obtain embeddings $\boldsymbol{\phi}_{\boldsymbol{\varepsilon}}^{\mathbf{n}}$, $\mathbf{f}_{\varepsilon}^{\mathbf{n}}$, $\boldsymbol{\phi}_{\boldsymbol{\varepsilon}}^{max}$ and $\mathbf{f}_{\varepsilon}^{\mathbf{max}}$:

$$w_{app}^n = \exp\left(\frac{\boldsymbol{\phi}_{\boldsymbol{\varepsilon}}^{\mathbf{n}} \cdot \boldsymbol{\phi}_{\boldsymbol{\varepsilon}}^{max}}{\|\boldsymbol{\phi}_{\boldsymbol{\varepsilon}}^{\mathbf{n}}\| \|\boldsymbol{\phi}_{\boldsymbol{\varepsilon}}^{max}\|}\right) \tag{6}$$

$$w_{flow}^n = \exp\left(\frac{\mathbf{f}_{\varepsilon}^{\mathbf{n}} \cdot \mathbf{f}_{\varepsilon}^{\mathbf{max}}}{\|\mathbf{f}_{\varepsilon}^{\mathbf{n}}\| \|\mathbf{f}_{\varepsilon}^{\mathbf{max}}\|}\right) \tag{7}$$

$$\boldsymbol{\phi}_{\mathbf{c}} = \sum_{t=1}^{n} w_{app^n} \boldsymbol{\phi}_{\mathbf{c}}^{\mathbf{n}} \tag{8}$$

A tracklet is a collection of person bounding boxes with same identity. Applying an average pooling operation on an embedding [16] appears like a sensible solution to summarize tracklet features. However, it is possible that one or more boxes in a tracklet could be occluded, or may contain other kinds of noise that may corrupt an average pooled tracklet. A tracklet can be compared to a cluster since a tracklet is a collection of same identities. Hence, applying the "max" operation on the tracklet is comparable to identifying the densest sample in a given cluster. The embedding that has maximum activation among all other embeddings can be assumed to have the most influence on the cluster center. Hence the Max embedding is used as a reference to calculate similarity with other embeddings in Eq. (6), allowing to calculate weights for each of embedding. These weights help determining the closeness of each embedding with the Max embedding (similar to a cluster center) which is the representative of a tracklet. Noisy embeddings that have lower activations can be disregarded by lower weights. The exponential operator ensures non-linearity, as well as assuring a minimum weight of 1 to each of the embedding. Thereby considering all the embeddings in a tracklet for the final representation without completely disregarding noisy features.

A similar approach is followed to accumulate optical flow embeddings with Eq. (7). In order to obtain one single embedding for a given tracklet, feature representations in the tracklet are aggregated by weighted addition assigning weights obtained in Eqs. (6) and (7) to each of the embeddings in a given tracklet. These weights can also be compared to attention-based weights where most important frames in a tracklet are given higher weights, and giving lower weights to noisy frames. From Eq. (8) we obtain outputs $\boldsymbol{\phi}_{\mathbf{c}}$ and when used with w_{flow} to obtain $\mathbf{f}_{\mathbf{c}}$ which are aggregated video features for image and flow respectively. These two features are concatenated to form $\boldsymbol{\phi}_{cat}$ which is passed though a fully connected layer of size same as $\boldsymbol{\phi}_{\mathbf{c}}$ to produce the final feature for classification or re-identification.

The network is trained on logistic loss and Triplet loss similar to the method used in [46]. During testing the fully connected layer is removed and the remainder of the network is used for feature extraction purpose and to match against those in gallery.

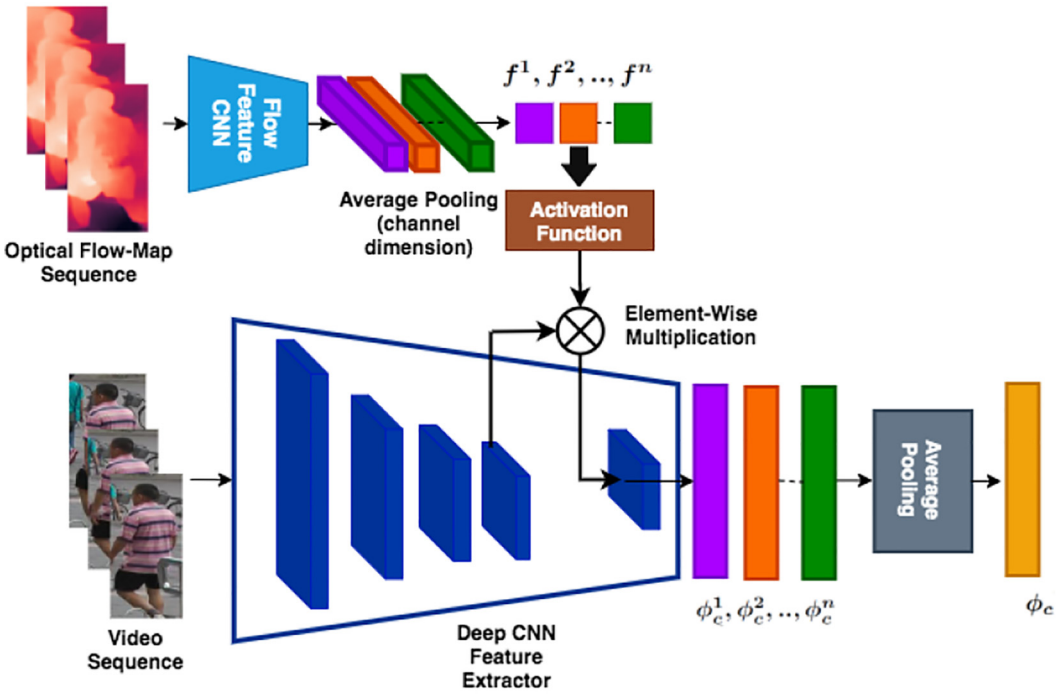

Fig. 4. Experimental setup for our baseline gated attention network. The system inputs a sequence of bounding boxed images and the corresponding optical flow maps from a given video clip. The features extracted from optical flow are pooled in channel dimension, and multiplied with intermediate layers of deep feature extraction after activation to obtain attended features for ReID. The network outputs a feature vector ϕ_c per video clip.

4. Experimental methodology

4.1. Datasets

Experiment are performed on 3 challenging and widely-used datasets for video-based person reID. MARS [61] dataset is one of the largest datasets for video Person-ReID. The dataset has been collected from six cameras with a total of 1261 identities. Another dataset commonly used in literature for evaluating video person ReID is Duke-MTMC [39,55] dataset containing 702 identities and more than 2000 sequences for testing and training each. Duke-MTMC dataset are of higher resolution compared to that of MARS. We also evaluate on ILIDS-VID [53] dataset, which has a total of 300 identities with videos across two cameras. The dataset is comprised of sequences from two disjoint camera views. One interesting point to be noted with ILIDS datasets is that the tracklets have been generated by hand annotation unlike detector based annotation in MARS dataset. This makes the bounding boxes well aligned in ILIDS-VID dataset enabling the optical flow estimation to be less noisy.

4.2. Settings

We follow the overall system architecture in [16] (baseline) and [46] (COSAM). They achieved state-of-the-art results on several ReID datasets using the ResNet50 CNN for feature extraction. We propose to use ResNet50 and SE-ResNet50 as our backbone networks to learn features invariant to cluttered background by attending with saliency map obtained from optical flow estimations. We have shown results with each of the networks as backbone separately. The networks have been pre-trained on the ImageNet [13] dataset. We experiment at different layers of the ResNet50 to select the ideal location in the network to generate maximum attention with optical flow. To extract video level feature from instance level features, we compare our proposed weighted addition method with that of temporal Average Pooling (AP) and Temporal Attention (TA) based method as illustrated in [16,46]. The Shallow CNN in our experimental setup for gated attention is based on AlexNet and the sub-Network for weighted addition is a two layer MLP of size 2048 nodes in each layer.

Common data augmentation methods such as random flips and random crops are followed during training. We use ADAM optimizer to train our model with a batch size of 32. We use a sequence length of 4 to train our model. The flow guided attention has been applied at layer 4 of the ResNet50 and an empirical study of attention at different

Table 1
Accuracy of our Baseline (ResNet50) with Temporal Pooling (TP) and our Baseline with Mutual Attention (MA) + TP on MARS dataset, over different layers of ResNet50.

Method	maP	Rank-1
Baseline [16]	75.8	83.1
Baseline + MA (Layer2)	78.2	84.5
Baseline + MA (Layer3)	78.8	84.9
Baseline + MA (Layer4)	**80.0**	**86.6**
Baseline + MA (Layer5)	78.1	84.3
Baseline + MA (All)	51.2	66.1

Table 2
An ablation study of contribution of different module, i.e., our Gated Attention and our Mutual Attention on the baselines models, using the MARS dataset.

Method	Feature Aggregation	mAP	Rank 1
Baseline [16] (One Stream)	Pooling	75.8	83.1
No Attention (Two Stream)	Pooling	76.7	84.3
Gated Attention (One Stream)	Pooling	77.4	84.6
Mutual Attention (Two Stream)	Pooling	79.1	85.4
Baseline [16]	RNN	75.7	82.9
Baseline [16]	Temporal Attention	76.7	83.3
Mutual Attention (Mutual Attention) Shared param	Weighted Addition	76.8	84.1
Mutual Attention (Mutual Attention) Separate param	Weighted Addition	**80.0**	**86.6**

layers has been presented in the next section. Hence the training setting have mostly been kept similar to our baselines [16]. In order to explore the advantage of attention with optical flow, we propose a separate experimental setup with simple gated attention mechanism i.e. using optical flow to attend to image stream alone. This one stream approach, where only image features are used for classification while optical flow is just used as an attention mechanism as described below.

4.2.1. Optical flow estimation

To estimate optical flow maps for a given sequence, the LiteFlowNet [22] model has been chosen as it is computationally efficient compared to other DL models, while obtaining state of the art performance. We have used the official implementation from the authors to produce flow maps for MARS, Duke-MTMC, and ILIDS-VID datasets. Hence for a given sequence, we input pairs of image I^{t-1},I^{t} as input to the LiteFlowNet model to produce flow map F^{t-1}.

4.2.2. Baseline for mutual attention

Given input images, $\mathbf{I_c^1}$,$\mathbf{I_c^2}$, …,$\mathbf{I_c^n}$ and $\mathbf{F_c^1}$,$\mathbf{F_c^2}$, …,$\mathbf{F_c^n}$ flow maps for images of a sequence, we extract the features $\boldsymbol{\phi_c} = (\boldsymbol{\phi_c^1}, \boldsymbol{\phi_c^2}, \ldots, \boldsymbol{\phi_c^n})$ and $\mathbf{f_c} = (\mathbf{f_c^1}, \mathbf{f_c^2}, \ldots, \mathbf{f_c^n})$ from the deep CNN and the Flow CNN respectively. A shallow CNN (Flow CNN) has been used to extract features from optical flow maps. We rely on shallow CNN to retain spatial coherence in the flow features (see Fig. 4).

Let l be the intermediate layer of the k layer deep CNN and let the deep CNN be represented by H with a total number of k layers. Let S denote the shallow flow feature extractor CNN with $t = 1,2,3\ldots n$ then:

$$\boldsymbol{\phi_c^t} = \mathbf{H}(\mathbf{I_c^t}) \ , \ \mathbf{f_c^t} = S(\mathbf{F_c^t}) \tag{9}$$

If features from layer l are expressed as ψ, then:

$$\boldsymbol{\psi_c^t} = H_l(\mathbf{I_c^t}) \tag{10}$$

The features from Eq. (10) are then pooled in the channel dimension to produce $N \times 1 \times I \times J$ feature for attention in the spatial dimension. The features are then activated by an activation function to produce a spatial soft attention map($A(f_c^t)$). where A is the sigmoid activation function, and a_c^t is the output of the activation function. Finally attention is applied to the intermediate features ψ at an intermediate layer to obtain activated features Ψ by element-wise multiplication with the activation a_c^t.

$$\Psi_\mathbf{c}^\mathbf{t} = \boldsymbol{\psi_c^t} \odot \mathbf{a_c^t} \tag{11}$$

After gated attention of features at the intermediate CNN layers, the feature extraction process is continued with the rest of the CNN layers. In Eq. (12) l, k represents layers between intermediate layer l and last layer k. Then, the output ϕ_c^t of feature extraction is given by,

$$\boldsymbol{\phi_c^t} = H_{l,k}(\Psi_\mathbf{c}^\mathbf{t}) \tag{12}$$

ϕ_c^t is further average pooled in temporal dimension to produce CNN features for re-identification. The network is trained with classification layer similar to our Mutual Attention Network. The gated attention network described here served as one of the baselines for flow guided attention.

4.3. Evaluation measures

During the training phase we learn the ReID task by training a classifier with identity labels from the single feature extracted for a sequence. The feature tractor produces a 2048 × 1 size feature vector per sequence. This is the input to train the ReID classifier. During the testing phase, we use the 2048 dimensional feature to measure distance between the test sequence and the sequence from the gallery. We use the Cumulative Matching Characteristic (CMC) and Mean Average Precision(mAP) to evaluate the performance. CMC represents the matching characteristics of the first n query results.

5. Results and discussion

First, this section provides an ablation study the assess the contribution of different modules to the performance of our proposed model. This study is performed on a systems using the baseline ResNet50 CNN architecture [16], using temporal pooling for feature aggregation. We also include an empirical study that allows selecting the deep feature extraction layer for embedding our attention mechanism, as well as the sequence length. The overall performance is compared against the baselines and state-of-the-art methods. We compare the overall performance on MARS, Duke-MTMC, and ILIDS-Vid dataset.

5.1. Flow guided attention fusion

The first part of our work consists of flow guided attention on the intermediate layer of the Deep CNN used for feature extraction in ReID. Our proposed network consists of flow guided attention on an intermediate layer of the Deep CNN used for feature extraction. Different layers in the Deep CNN have different abstraction level for salient features of the input person image. Hence an experiment was conducted by fusing the flow guided attention at different layers of the Depp CNN applied on the baseline [16] ReID system, and evaluated on MARS dataset. From the Table 1 we conclude that the best performance was achieved by attending at layer 4 of the ResNet50 network. This is justifiable from the fact that earlier layers have different abstraction level for salient features – the level of abstraction increases in the later layers but, in the last layer, the spatial coherence is lower than in previous layers. We also conduct an experiment where we apply our MA to multiple layers of ResNet50 (namely to layers 2, 3, 4 and 5), and obtained the results shown in Table 1. It can be observed that multi-layer attention provides much lower accuracy than expected. This may be due to the attention from optical flow being sparse, and thus applying them on all layers drastically reduces the overall number of neurons activated across the CNN.

Table 3
Accuracy of our proposed Mutual Attention (MA), and baselines with no attention, with average pooling, and with RNN for different video sequence length on MARS dataset.

No of frames per sequence	Baseline [16]		RNN Aggregation		Ours			
	Average Pooling No Attention				Gated Attention Weighted Addition		Mutual Attention Weighted Addition	
	MAP	Rank 1	mAP	Rank 1	mAP	Rank 1	mAP	Rank 1
2	71.0	81.8	–	–	77.0	84.0	74.8	82.4
4	75.1	83.2	75.7	82.9	77.8	84.8	77.7	85.4
6	74.4	82.7	–	–	77.6	84.5	79.2	85.8
8	73.3	82.0	76.2	82.5	77.3	84.2	79.3	86.4
16	–	–			72.5	82.9	**80.0**	**86.6**

Table 4
Accuracy of our proposed vs state-of-the-art methods evaluated on the MARS and Duke-MTMC datasets. Column "Opt. Flow" refers to the use of optical flow as additional inputs.

Method	Opt Flow	Reference	MARS		Duke-MTMC	
			mAP	Rank-1	mAP	Rank-1
LOMO + SQDA [28] Histogram based	No	CVPR 2015	16.4	30.7	–	–
Set2set [31] Custom Network	No	CVPR 2017	51.7	73.7	–	–
JST RNN [64] CaffeNet	No	CVPR 2017	50.7	70.6	–	–
k-reciprocal [63] ResNet50	No	CVPR-2017	58.0	67.8	–	–
TriNet [18] ResNet50	No	ArXiv 2019	67.7	79.8	–	–
RQEN [44] GoogLeNet	No	AAAI 2018	71.1	**77.8**	–	–
Part-Alignment [47] GoogLeNet	No	ECCV 2018	72.2	83.0	78.3	83.6
Mask-Guided [43] ResNet50	No	CVPR 2018	71.1	77.1	–	–
Snippet [6] ResNet50	Yes	CVPR 2018	71.1	82.1	–	–
STA [15] ResNet50	No	AAAI-2019	80.8	86.3	94.9	96.2
RevstTemPool [16] ResNet50	No	Arxv 2018	75.8	83.1	–	–
Cosam [46] ResNet50	No	ICCV 2019	76.9	83.6	93.5	93.7
STAL [7] ResNet-50	No	IEEE TIP 2019	73.5	82.2	–	–
Cosam [46] SE-ResNet50	No	ICCV 2019	79.9	84.9	94.1	95.4
STAR [54] ResNet-50		Yes	BMVC 2019	76.0	85.4	–
SCAN [57] ResNet-50	No	IEEE TIP 2019	76.7	86.6	–	–
SCAN [57] ResNet-50	Yes	IEEE TIP 2019	77.2	87.2	–	–
GLTR [25] ResNet-50	Yes	CVPR 2019	78.5	87	93.7	96.3
M3D [27] ResNet-50	No	IEEE TIP 2020	79.46	88.63	93.67	95.49
Rec3D [8] ResNet-50	Yes	IEEE TIP 2020	80.4	86.3	–	–
MultiGrain [58] ResNet-50	No	CVPR 2020	**85.9**	**88.8**	–	–
Mutual Attention ResNet-50	Yes	**Ours**	80.0	86.6	**94.9**	95.6
Mutual Attention SE-ResNet-50	Yes	**Ours**	80.9	87.3	94.8	**96.7**

Table 5
Accuracy of our proposed vs state-of-the-art methods evaluated on the ILIDS-Vid dataset. "Opt. Flow" refers to the use of optical flow as additional inputs.

Method	Opt. Flow	Reference	Rank 1	Rank 5
Two Stream* [12]	Yes	ICCV 2017	60.0	86.0
Snippet [6] ResNet50	Yes	CVPR-2018	85.4	87.8
RQEN [44] GoogLeNet	No	AAAI-2018	77.1	93.2
STAL [7] ResNet-50	No	IEEE TIP	82.8	95.3
COSAM SE-ResNet50 [46]	No	ICCV-2019	79.6	95.3
GLTR [25] ResNet-50	Yes	CVPR-2019	86.0	–
Rec3D [8] ResNet-50	Yes	IEEE TIP-2020	87.9	98.6
Mutual Attention ResNet-50	Yes	Ours	86.2	96.4
Mutual Attention SE ResNet-50	Yes	Ours	**88.1**	98.4

Table 6
A comparison of attention methods and time complexity (in Flops) of some SOA methods.

Model	Attention Method	Optical Flow	Time Complexity
Mask Guided [43]	Video based + binary segmented mask.	No	N/A
COSAM [46]	Video co-segmentation based	No	35 G
GLTR [25]	Video Based temporal self attn	Yes	60 G
SCAN [57]	Video based inter-sequence attn	Yes	70 G
STAR [54]	Video frame level association based	Yes	N/A
REC3D [8]	3D attn. by Reinforcement Learning	Yes	N/A
Snippet [6]	Video LSTM based aggregation	Yes	60 G
MA (Ours)	Two stream mutual attention	Yes	58 G

5.2. Contribution of different modules to the baseline

In this subsection we compare Mutual Attention methods to the baseline [16] since we follow similar overall system architecture. In Table 2 we compare the baseline and ours with different feature aggregation methods like Average Pooling, Temporal Attention and our weighted feature addition method described earlier. We also compare our Mutual Attention method with single stream with Gated Attention using optical flow described in the introduction to Experiments section. We can see that just Gated Attention on its own on the baseline [16] has improved the performances by a large margin on MARS datasets. Our Mutual Attention method further improved the results compared to Gated Attention showing the potential of Mutual Attention between both image and optical flow features. Out Feature addition method used with Mutual Attention improve the results for feature aggregation by a larger margin compared to both Average Pooling and aggregation method from Gated Attention. Table 2 has additional results with an ablation study for using shared backbone parameters for the dual stream. It can be observed from the table that using shared parameters ("shared param" in table) between optical flow network and ResNet produces much lower results than using separate parameters ("separate param" in table). This is because of the deference in the representation space of both optical flow and image. Although optical flow resembles a silhouette of the person image, it lacks many fine grained spatial features such as intricate shapes and textures that can be seen in the image space. CNN backbone filters of image based stream tend to learn the different shape and texture information which might not be suitable to learn optical flow information and vice versa. Thereby using separate CNN backbones would enable each of the backbones to learn ideal filters for the corresponding input streams.

5.3. Effect of sequence length

The length of the sequence has an effect on the representative power of final aggregated feature. This in-turn influences the performance of various feature aggregation methods. Therefore in this subsection we analyze the effect of sequence length on different feature aggregation methods such as Temporal Pooling, Flow guided weighted Addition applied on our method. Hence in Table 3 we have shown results of flow guided attention on ResNet50 architecture with both the feature aggregation methods. It can be seen that at the sequence length of 4 we obtain ideal results for most methods in the literature as well as for the simple gated attention method. But our Mutual Attention method demonstrate the ability to aggregate additional features and hence we could use a sequence of length 16 with Mutual Attention. This is a crucial result as we demonstrate ability to aggregate additional features and keep improving results until a sequence length of 16. Longer sequences have attributed to long term better motion and appearance features. At the same time in other methods in literature, simple averaging adds

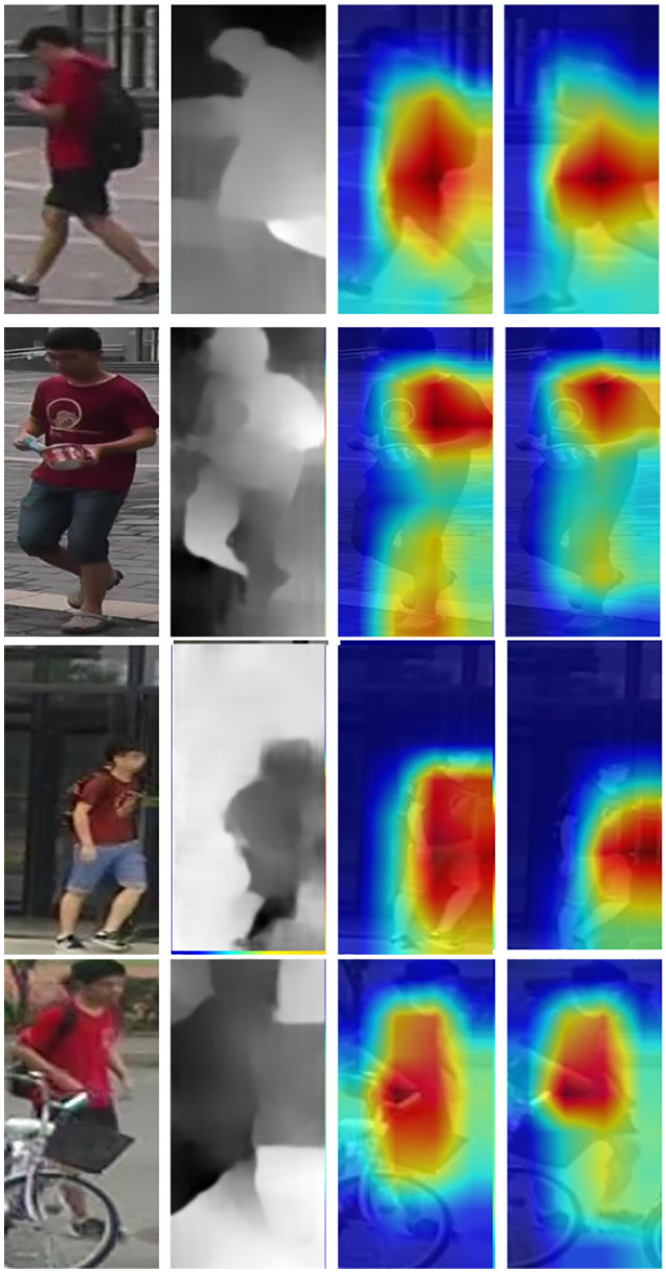

Fig. 5. Visualization of feature maps produced using bounding box images from the MARS dataset. The first column shows the original input image, the second column shows corresponding optical flow, and the third column shows the corresponding activation maps of our proposed attention. The fourth column shows the activation map for our baseline, generated by training using our backbone without using the optical flow and keeping the rest of the setting same.

additional noise to the features with longer sequences. Our weighted addition methods weights the individual feature based on importance and relevance thereby reducing noise with longer sequences.

5.4. Comparison with state-of-the-art

We report the performance of our method with backbones ResNet50 and SE-ResNet50 [21] separately with our Mutual Attention and weighted feature aggregation method on MARS, Duke-MTMC datasets and ILIDS-VID [53] in the Tables 4 and 5 compared with related works. As mentioned earlier we have selected [16] as our baseline. It can be observed that from our baseline. we have improved by a large margin on both mAP and Rank1 metric. Our method has also outperformed most of the state-f-the-art methods including some of the best existing methods. We have also shown the advantage of our method compared to other optical flow based methods [6,54,57]. Although [29] have demonstrated State-Of-the-Art results, we do not compare with them as their evaluation strategy is different from that of the commonly followed method in literature. We attribute our performance gain compared to the baseline on both flow guided attention and our feature aggregation technique. It can also be observed that from our proposal Mutual Attention method performs best demonstrating that optical flow and image stream can attend to the salient regions of each other. Our proposed Mutual Attention method could be integrated with any back-end architecture from some of the strong baselines to achieve better performance. Since our main aim was to evaluate the contribution of Mutual Attention clearly, we choose a simple baseline [16]. Compared to the results on other datasets, the margin of improvement is higher with ILIDS-VID dataset due uniformly centered hand-annotated person cutouts of ILIDS-VID dataset. Results also suggest that our proposed approach is vulnerable to quality of person detection and tracking.

We compare the complexity and attention method used in recent SOA methods along with their complexities in GFLOPs (See Table 6). It can be observed that the complexity of our method is somewhat close to the other SOA methods although COSAM [46] is more efficient comparatively.

5.5. Visualization

Fig. 5 shows activations of feature maps from final layer of our backbone 2D-CNN feature extractor on some bounding box images from the MARS dataset. The first columns shows original input image, the second column shows the corresponding optical flow, the third column shows activation after our proposed attention. The fourth column shows the activation map of our baseline that was generated by training our backbone without relying on optical flow, but that preserves the other settings. This was chosen as our baseline because the SOA methods use different backbones and different experimental settings. Our baseline is similar to [16]. It can be observed from the examples shown in Fig. 5 that our proposed flow guided mutual attention enhanced the spatial activations of feature maps based on the optical flow produced by the virtue of motion of persons between different consecutive frames of the sequence. Using the proposed approach, some additional regions of the person are activated (corresponding to moving parts), while the baseline approach only focuses mostly on a potentially discriminant spatial region.

It can be observed that the overall noise from the background has been suppressed particularly in the last row where the person is poorly localized. Fig. 6 shows a tracklet with its activations and the last row shows single tracklet feature 1) by proposed weighted addition and 2) average pooled. It can be observed that our proposed weighted addition method captures activation well from salient regions in the tracklet (Shows a good response for both head and legs compared to average pooled version).

6. Conclusion

In this work we present a novel framework for flow guided attention and temporal feature aggregation for Person-ReID. We focus on visual appearances across spatial and temporal streams and their correlations to encode common saliency between the streams, reduce background clutter, learn motion patterns of person and to have the advantages of having longer sequences. Our feature aggregation method uses cues from both image and optical flow feature to assign weights and aggregate image instance features to produce a single video feature representation unlike assigning equal weights to images instances as in temporal pooling. Our method outperforms the state-of-the-art person reID

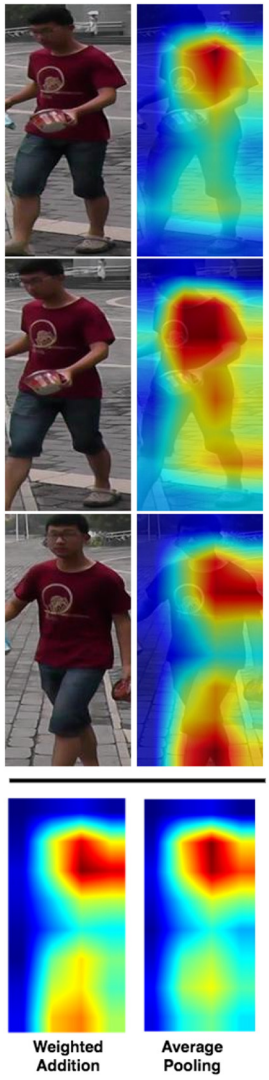

Fig. 6. Visualization of features of a tracklet from the MARS dataset and our proposed weighted addition of the features vs average pooling of the features to extract one aggregated feature for the tracklet. – The last row shows our proposed weighted addition for the features of the tracklet, and the average pooled version of the tracklet above it, respectively.

methods in terms of both mAP and Rank1 accuracy evaluated on MARS, Duke-MTMC, and ILIDS-VID datasets. The proposed Mutual Attention network is most effective when the person detection and tracking produces high quality bounding-boxes, and in scenarios with bigger-sized bounding boxes for objects, where the attention helps in locating the objects.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was partially supported by the Mathematics of Information Technology and Complex Systems (MITACS) and the Natural Sciences and Engineering Research Council of Canada (NSERC) organizations.

References

[1] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, CVPR, 2015.
[2] A. Bhuiyan, Y. Liu, P. Siva, M. Javan, I.B. Ayed, E. Granger, Pose guided gated fusion for person re-identification, WACV, 2020.
[3] A. Bhuiyan, A. Perina, V. Murino, Person re-identification by discriminatively selecting parts and features, ECCV, 2014.
[4] A. Bhuiyan, A. Perina, V. Murino, Exploiting multiple detections for person re-identification, J. Imag. 4 (2018) 28.
[5] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a "siamese" time delay neural network, NIPS, 1994.
[6] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang, Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding, ICCV 2018, pp. 1169–1178, https://doi.org/10.1109/CVPR.2018.00128.
[7] G. Chen, J. Lu, M. Yang, J. Zhou, Spatial-temporal attention-aware learning for video-based person re-identification, IEEE Trans. Image Process. 28 (2019) 4192–4205.
[8] G. Chen, J. Lu, M. Yang, J. Zhou, Learning recurrent 3d attention for video-based person re-identification, IEEE Trans. Image Process. 29 (2020) 6963–6976.
[9] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, CVPR, 2017.
[10] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, CVPR, 2016.
[11] S. Cho, H. Foroosh, Spatio-temporal fusion networks for action recognition, ACCV, Springer 2018, pp. 347–364.
[12] D. Chung, K. Tahboub, E.J. Delp, A two stream siamese convolutional neural network for person re-identification, ICCV 2017, pp. 1992–2000, https://doi.org/10.1109/ICCV.2017.218.
[13] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, CVPR09, 2009.
[14] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, CVPR, 2010.
[15] Y. Fu, X. Wang, Y. Wei, T. Huang, Sta: Spatial-temporal attention for large-scale video-based person re-identification, AAAI, 2019, 2019, pp. 8287–8294.
[16] J. Gao, R. Nevatia, Revisiting temporal modeling for video-based person reid. arXiv preprint, arXiv:1805.02104 2018.
[17] X. Gu, B. Ma, H. Chang, S. Shan, X. Chen, Temporal knowledge propagation for image-to-video person re-identification, ICCV, 2019.
[18] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification. arXiv preprint, arXiv:1703.07737 2017.
[19] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Vrstc: Occlusion-free video person re-identification, CVPR, 2019.
[20] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, CVPR 2018, pp. 7132–7141, https://doi.org/10.1109/CVPR.2018.00745.
[21] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 7132–7141.
[22] T.W. Hui, X. Tang, C.C. Loy, Liteflownet: a lightweight convolutional neural network for optical flow estimation, CVPR 2018, pp. 8981–8989.
[23] M. Kiran, R.G. Praveen, L.T. Nguyen-Meidine, S. Belharbi, L.A. Blais-Morin, E. Granger, Holistic guidance for occluded person re-identification. arXiv preprint, arXiv:2104.06524 2021.
[24] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, CVPR, 2017.
[25] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-local temporal representations for video person re-identification, CVPR 2019, pp. 3958–3967.
[26] J. Li, S. Zhang, T. Huang, Multi-scale 3d convolution network for video based person re-identification, AAAI 2019, pp. 8618–8625.
[27] J. Li, S. Zhang, T. Huang, Multi-scale temporal cues learning for video person re-identification, IEEE Trans. Image Process. 29 (2020) 4461–4473.
[28] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, CVPR, 2015.
[29] C.T. Liu, C.W. Wu, Y.C.F. Wang, S.Y. Chien, Spatially and temporally efficient non-local attention network for video-based person re-identification, BMVC, 2019.
[30] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification, IEEE Trans. Image Process. 26 (2017) 3492–3506.
[31] Y. Liu, Y. Junjie, W. Ouyang, Quality aware network for set to set recognition, CVPR, 2017.
[32] C.Y. Ma, M.H. Chen, Z. Kira, G. AlRegib, Ts-lstm and temporal-inception: exploiting spatiotemporal dynamics for activity recognition, Signal Process. Image Commun. 71 (2019) 76–87.
[34] N.D. McLaughlin, J.M. Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, CVPR 2016, pp. 1325–1334, https://doi.org/10.1109/CVPR.2016.148.
[35] D. Mekhazni, A. Bhuiyan, G. Ekladious, E. Granger, Unsupervised domain adaptation in the dissimilarity space for person re-identification, European Conference on Computer Vision, Springer 2020, pp. 159–174.
[36] R. Panda, A. Bhuiyan, V. Murino, A.K. Roy-Chowdhury, Unsupervised adaptive re-identification in open world dynamic camera networks, CVPR, 2017.
[37] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, ECCV, 2018.
[38] R. Quan, X. Dong, Y. Wu, L. Zhu, Y. Yang, Auto-Reid: Searching for a Part-Aware Convnet for Person Re-Identification, 2019.
[39] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, ECCVWK, 2016.

[40] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, ICCV, 2018.
[41] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A.C. Kot, G. Wang, Dual attention matching network for context-aware feature sequence based person re-identification, CVPR 2018, pp. 5363–5372, https://doi.org/10.1109/CVPR.2018.00562.
[42] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in Neural Information Processing Systems 2014, pp. 568–576.
[43] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, CVPR 2018, pp. 1179–1188, https://doi.org/10.1109/CVPR.2018.00129.
[44] G. Song, B. Leng, Y. Liu, C. Hetang, S. Cai, Region-based quality estimation network for large-scale person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[45] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, ICCV, 2017.
[46] A. Subramaniam, A. Nambiar, A. Mittal, Co-segmentation inspired attention networks for video-based person re-identification, ICCV, 2019.
[47] Y. Suh, J. Wang, S. Tang, T. Mei, K.M. Lee, Part-aligned bilinear representations for person re-identification, ECCV, 2018.
[49] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), ECCV, 2018.
[50] C.P. Tay, S. Roy, K.H. Yap, Aanet: Attribute attention network for person re-identifications, CVPR, 2019.
[51] P. Turaga, R. Chellappa, A. Veeraraghavan, Advances in video-based human activity analysis: challenges and approaches, in: M.V. Zelkowitz (Ed.), Advances in Computers, Elsevier 2010, pp. 237–290 , volume 80 of *Advances in Computers*. URL: http://www.sciencedirect.com/science/article/ pii/S0065245810800075, doi:doi:10.1016/S0065-2458(10) 80007-5.
[52] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, ECCV, 2016.
[53] T. Wang, S. Gong, X. Zhu, S. Wang, Person re-identification by video ranking, ECCV, 688–703, Springer, 2014.
[54] G. Wu, X. Zhu, S. Gong, Spatio-temporal associative representation for video person re-identification, BMVC 2019, p. 278.
[55] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, 2018 CVPR.
[56] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, ICPR, 2014.
[57] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, L. Lin, Scan: self-and-collaborative attention network for video person re-identification, IEEE Trans. Image Process. 28 (2019) 4870–4882.
[58] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 10407–10416.
[59] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: person re-identification with human body region guided feature decomposition and fusion, CVPR, 2017.
[60] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, CVPR, 2017.
[61] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, Mars: a video benchmark for large-scale person re-identification, European Conference on Computer Vision, Springer 2016, pp. 868–884.
[62] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose invariant embedding for deep person re-identification. arXiv preprint, arXiv:1701.07732 2017.
[63] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, CVPR, 2017.
[64] Z. Zhou, Y. Huang, W. Wang, L. Wang, T. Tan, See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification, CVPR 2017, pp. 6776–6785, https://doi.org/10.1109/CVPR.2017.717.