Flow-Guided Attention Networks for Video-Based Person Re-Identification

Madhu Kiran\textsuperscript{a}, Amran Bhuiyan\textsuperscript{a}, Louis-Antoine Blais-Morin\textsuperscript{b}, Mehrsan Javan\textsuperscript{c}, Ismail Ben Ayed\textsuperscript{a}, Eric Granger\textsuperscript{a}

\textsuperscript{a} Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Montreal, Canada
\textsuperscript{b} Genetec Inc., Montreal Canada
\textsuperscript{c} Sportlogiq Inc., Montreal Canada

mkiran@livia.etsmtl.ca, amran.apece@gmail.com
lablaismorn@genetec.com, mehrsan@gmail.com
Ismail.BenAyed@etsmtl.ca, eric.granger@etsmtl.ca

Abstract

Person Re-Identification (ReID) is an important problem in many video analytics and surveillance applications, where a person’s identity must be associated across a distributed network of cameras. Video-based person ReID has recently gained much interest because it can capture discriminant spatio-temporal information that is unavailable for image-based ReID. Despite recent advances, deep learning 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, two different flow-guided attention networks are proposed for fusion with any 2D-CNN backbone, allowing to encode temporal information along with spatial appearance information.

1. Introduction

Person Re-Identification (ReID) refers to the problem of associating individuals over a set of non-overlapping camera views. It is one of the key object recognition tasks, that has recently drawn a significant attention due to its wide range of applications in monitoring and surveillance, 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., illumination, blur), occlusions and background clutter.

ReID approaches can be applied in image-based and video-based settings. State-of-the-art\cite{1,2,40,29,41,11,3,4,27} approaches on image-based setting seek to associate still images of individuals over a set of non-overlapping cameras. More recently, the advances in computational resources have allowed researchers focus on addressing this problem in video-based setting, where input video tracklet\footnote{A tracklet correspond 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.} of an individual are matched against a gallery of tracklet representations, captured with different non-overlapping cameras. Compared to image data, video data provides richer source of information about persons’ appearance along with motion information that notably capture persons’ body layouts. Thus, video-based approaches allow to exploit spatio-temporal information (appearance and motion) for discriminative feature representation.

As illustrated in Figure\footnote{State-of-the-art approaches for video-based person ReID typically learn global features in an end-to-end fashion, through various temporal feature ag-} state-of-the-art approaches for video-based person ReID typically learn global features in an end-to-end fashion, through various temporal feature ag-
In this paper, we focus on using the image-level semantic information obtained from optical flow as a source of spatial attention for RGB images in a video clip. Since an optical flow map represents the change between two consecutive images of a moving object, it can serve as a cue for weighing instance-level image features for feature aggregation based on quality of semantic information in a given image. This has not been explored in the literature for video person-ReID, thereby undermining the global saliency in the feature representation by using optical flow for discriminative feature extraction. We focus on generating spatio-temporal attention based on optical flow such that the salient features from optical flow allow focusing on semantic information at the instance level (spatial attention), as well as feature aggregation at the sequence level (temporal attention), along with learning motion and appearance based matching cues for ReID.

Two new DL models for flow-guided attention are introduced in this paper – which we refer to as Gated Attention and Mutual Attention networks – for video-based person ReID with spatio-temporal attention, using any 2D-CNN backbone. They enabled to co-jointly learn a feature embedding that incorporates relevant spatial information from human appearance, along with their motion information, which we hereafter refer as Gated Attention and Mutual At-

### Figure 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 image. The features are then aggregated to produce a single feature representation for the clip, which is then matched against clip representations stored in the gallery.

### Figure 2
Example of sampling strategy for training or query a ReID system. Longer clips capture a wider range of variations, and while shorter clips are sensitive to issues like occlusion, lighting changes, poor bounding box alignment, etc.
2. Related Work

The section provides background on DL models for spatio-temporal recognition, optical flow, and attention networks.

The proposed Gated Attention network consists of two parallel streams, an appearance learning stream (backbone CNN) and a flow feature extractor CNN (Flow-CNN) which serves as an auxiliary supplement to the backbone net for robust person ReID. The Flow-CNN extracts features from optical flow to produce gated attention that attends the features of appearance learning stream which is in contrast with earlier works on optical flow discussed above. The Gated Attention network uses optical flow strictly for attention and to identify salient temporal features during aggregation. The Mutual Attention network includes both optical flow stream and image stream for ReID and leverages the joint information. We also demonstrate one feature aggregation method with each of the models.

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 using both optical flow and image feature information. During feature aggregation, a reference frame from each tracklet is selected based on maximum activation from both the streams, and weights are assigned for individual features using image and flow feature information. Attention is enabled from optical flow in both spatial and temporal domain to extract discriminant features for ReID.

Performance of the proposed flow-gated attention networks are evaluated and compared on the challenging MARS and Duke-MTMC datasets for video-based person ReID. Experimental results show that both improve accuracy and can outperform the state-of-the-art approaches. Results also indicate the capability for higher accurate predictions by using longer video clips to capture multiple appearance variations.

2. Related Work

The section provides background on DL models for spatio-temporal recognition, optical flow, and attention mechanisms as they relate to person ReID.

a) Image-Based Person-ReID: Person-ReID for still images has made significant strides in recent years, especially after the introduction of DL networks. The idea of using CNNs for re-ID 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 [46] 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, 8, 43, 7, 23, 3] 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, 43], 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 [36, 38, 50, 48, 28, 49, 31, 40] rely on incorporating contextual information into the base deep ReID model, where local and global feature representations are combined to improve accuracy. A few attention-based approaches for deep re-ID [19, 49, 36] 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 [40] or by pose-based part representation [36, 38, 50, 48, 28, 49, 31]. Although these methods have have achieve considerable performance improvements, they fail to incorporate temporal information due to their image-based setting.

b) Video-Based Person-ReID: Video Person-ReID has become prevalent in practical applications, with the potential advantage of exploiting temporal information along with spatial information to improve performance. Video ReID has recently attracted some interest since temporal information allows dealing with ambiguities such as occlusion and background noise [12, 13, 37, 15, 25]. 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. [12] 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 normalised 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 [12]. 3DCNN has been experimented in [12, 20] but have not been very effective in summarising 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.

c) 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 [16]. Attention mechanisms for person ReID have been proposed for self-attention or guided by some additional information outside a network. A dual attention mechanism has also been proposed in [32] to perform intra-sequence feature refinement and inter-sequence feature pair alignment. This allows to exploit visual cues in intermediate frames. A mask guided attention mechanism has been proposed in [34], where a binary body mask is used in conjunction with the corresponding person image to reduce background clutter. Somewhat similar to [34], 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 [37]. Co-Segmentation allows extracting common saliency between images, and using this information for both spatial- and channel-wise attention.

Similar to co-segmentation, optical flow is another important cue for common saliency between objects in motion captured between two consecutive images in a sequence. Optical flow is defined as the apparent motion of individual pixels on the image plane. It often serves as a good approximation of the true physical motion projected onto the image plane [42]. Optical flow has been employed for temporal information fusion in [6, 20], 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 [9, 33] 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 [9]. Therefore there is a necessity to use optical flow in a way that it can be leveraged for appearance related task.

It can be summarised from the above that, as discussed in [37], [32], [34] and [12], various saliency feature enhancement methods have been attempted, and in most cases they have helped improve the overall performance. Optical flow typically encodes motion information in contrast to appearance information, and hence there is scope to explore enhancing appearance information from motion and vice-versa.

3. Proposed Methods

Given an input video clip (set of bounding boxes extracted from a tracklet) represented by \( I^1_c, I^2_c, ..., I^n_c \) and corresponding optical flow estimations \( F^1_c, F^2_c, ..., F^n_c \), where \( c \) indicates the Identity of the video clip of length \( n \), our objective is to extract a discriminative feature vector \( \phi_c \) for ReID.

Two new models are proposed for flow guided attention – the Gated Attention and Mutual Attention networks. They learn spatio-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. For temporal attention, they perform weighted addition of features of individual image frames by using attention that is a generated based on image and flow features. The Gated Attention network is a simple gated attention model, where the optical flow stream is used strictly for attention, and not directly as an input streams for matching. The Mutual Attention network includes both optical flow and image streams to attend each other, and also to combine the features to yield a single feature representation per clip.

3.1. Gated Attention Network:

**Flow Guided Attention.** Given input images, \( I^1_c, I^2_c, ..., I^n_c \) and flow maps \( F^1_c, F^2_c, ..., F^n_c \) for images of a sequence, we extract the features \( \phi_c = (\phi^1_c, \phi^2_c, ..., \phi^n_c) \) and \( f_c = f^1_c, f^2_c, ..., f^n_c \) from the deep CNN and the Flow CNN respectively. Since deep CNNs have been effective in Person-ReID problems, it is used for feature extraction in the Gated Attention network. 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).

The saliency level of the image feature varies across dif-
determined by the spatial dimension. The features are then activated by an activation function to produce a spatial soft attention map.

\[ a^c_t = A(f^c_t) \]  

where \( A \) is the sigmoid activation function, and \( a^c_t \) is the output of the activation function of size \( N \times 1 \times I \times J \). Finally, attention is applied to the intermediate features \( \Psi \) at an intermediate layer to obtain activated features \( \psi^c_t \) by element-wise multiplication with the activation \( a^c_t \).

\[ \psi^c_t = \psi^c_t \odot a^c_t \]  

A sequence of flow estimations input into the network will produce a sequence of flow features of size \( N \times C \times I \times J \), representing sequence length, channels, width, height respectively. The features 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.

\[ \phi^c_t = H(I^c_t) \]  

where \( H \) is the sigmoid activation function, and \( \phi^c_t \) is the output of the activation function of size \( N \times 1 \times I \times J \). This is not a fair weighing strategy as it implies that the instance level features are weighted equally and summed. This is not a fair weighing strategy particularly if one of the instance is partially occluded or cluttered. Temporal attention strategy is similar to our proposed method where a set of network parameters generate attention weights to aggregate the features except that they rely on image features alone.

We propose to use feature weighing strategy proposed by [18] for attention based feature aggregation. By considering both flow feature \( f^c_t \) as well as image feature \( \phi^c_t \) for different layers of the CNN. The flow guided attention is fused at a specific layer in the deep CNN where the attention is maximum. 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 \( l = 1, 2, 3...n \) then,

\[ \phi^c_t = H(I^c_t) \]  

If features from layer \( l \) are expressed as \( \psi \), then

\[ \psi^c_t = H(I^c_t) \]  

A sequence of flow estimations input into the network will produce a sequence of flow features of size \( N \times C \times I \times J \), representing sequence length, channels, width, height respectively. The features 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.

\[ a^c_t = A(f^c_t) \]  

where \( A \) is the sigmoid activation function, and \( a^c_t \) is the output of the activation function of size \( N \times 1 \times I \times J \). Finally, attention is applied to the intermediate features \( \Psi \) at an intermediate layer to obtain activated features \( \psi^c_t \) by element-wise multiplication with the activation \( a^c_t \).

\[ \psi^c_t = \psi^c_t \odot a^c_t \]  

Weighted Feature Addition. To generate one single feature vector \( \phi^c_c \) for a sequence \( c \) from instances \( \phi^c_1, \phi^c_2, ..., \phi^c_n \), various methods have been described in [12] including methods based on RNNs, average pooling and temporal attention. As described earlier, average pooling would imply that the instance level features are weighed equally and summed. This is not a fair weighing strategy particularly if one of the instance is partially occluded or cluttered. Temporal attention strategy is similar to our proposed method where a set of network parameters generate attention weights to aggregate the features except that they rely on image features alone.

We propose to use feature weighing strategy proposed by [18] for attention based feature aggregation. By considering both flow feature \( f^c_t \) as well as image feature \( \phi^c_t \) for different layers of the CNN. The flow guided attention is fused at a specific layer in the deep CNN where the attention is maximum. 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 \( l = 1, 2, 3...n \) then,

\[ \phi^c_t = H(I^c_t) \]  

If features from layer \( l \) are expressed as \( \psi \), then

\[ \psi^c_t = H(I^c_t) \]  

A sequence of flow estimations input into the network will produce a sequence of flow features of size \( N \times C \times I \times J \), representing sequence length, channels, width, height respectively. The features 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.

\[ a^c_t = A(f^c_t) \]  

where \( A \) is the sigmoid activation function, and \( a^c_t \) is the output of the activation function of size \( N \times 1 \times I \times J \). Finally, attention is applied to the intermediate features \( \Psi \) at an intermediate layer to obtain activated features \( \psi^c_t \) by element-wise multiplication with the activation \( a^c_t \).

\[ \psi^c_t = \psi^c_t \odot a^c_t \]  

After gated attention of features at the intermediate CNN layers, the feature extraction process continues with the rest of the CNN layers. If \( \Psi^c_t \) is the output of attention, it is passed through the rest of the layers in the network. In Eqn. 5 \( l, k \) represents layers between intermediate layer \( l \) and last layer \( k \). Then, the output \( \phi^c_t \) of feature extraction is given by,

\[ \phi^c_t = H_{l,k}(\psi^c_t) \]  

In Eqn. 6, \( w \) has been generated by a function of both image feature and flow feature:

\[ w^c_t = \exp \left\{ \frac{Z^T \left( \tanh \left( V \phi^c_t \odot U f^c_t \right) \right)}{\sum_{t=1}^{n} \exp \left\{ Z^T \left( \tanh \left( V \phi^c_t \odot U f^c_t \right) \right) \right\}} \right\} \]
is illustrated in Fig. 5 where the network accepts a two stream 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.

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

Given a video clip and corresponding flow maps, we extract the features $\phi_c$ and $f_c$ 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 $H_{\text{app}}$ and optical flow stream CNN by $H_{\text{flow}}$ with a total number of $k$ layers. With $t = 1, 2, 3...n$ we have:

$$
\phi_c^l = H_{\text{app}}(I_c^l), \quad f_c^l = H_{\text{flow}}(F_c^l) \quad (8)
$$

If features from layer $l$ are expressed as $\phi^l$, then

$$
\phi_c^{l, t} = H_{\text{app}}(I_c^l), \quad f_c^{l, t} = H_{\text{flow}}(F_c^l) \quad (9)
$$

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,

$$
correl = \zeta_{\text{app}}(\phi_c^{l, t}) \odot \zeta_{\text{flow}}(f_c^{l, t}) \quad (10)
$$

In the Eqn. 10 $\zeta_{\text{app}}$ and $\zeta_{\text{flow}}$ are the embeddings with $1x1$ convolution with ReLU discussed above. correl when activated by a sigmoid function forms the mutual attention map $M_c^l$ between both the streams of input.

$$
M_c^l = \text{Sigmoid}(\text{correl}) \quad (11)
$$

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

$$
\Psi_{\text{app}, c}^t = \phi_c^{l, t} \odot M_c^l, \quad \Psi_{\text{flow}, c}^t = f_c^{l, t} \odot M_c^l \quad (12)
$$

**Weighted Feature Addition.** We consider another method to aggregate image level features to obtain a single feature vector for a given video sequence. 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 $\phi_c$ for a sequence $c$ from instances $\phi_{c, 1}, \phi_{c, 2}, ..., \phi_{c, n}$ and $f_c$ from $f_{c, 1}, f_{c, 2}, ..., f_{c, n}$. The first task is to identify salient feature from a given sequence of features. In our case we can define a salient feature 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 as reference frame denoted by $\phi_{c, \text{max}}$ and $f_{c, \text{max}}$. 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 $\phi_{c, n}$ and $f_{c, n}$. Instead a tiny embedding $\epsilon(\cdot)$ is applied on the $\phi_{c, n}^\text{app}, f_{c, n}^\text{flow}$ and reference feature $\phi_{c, \text{max}}^\text{app}, f_{c, \text{max}}^\text{flow}$ to obtain embeddings $\phi_{\text{app}}^\epsilon, f_{\text{flow}}^\epsilon$ which are aggregated video features for image and flow respectively. These two features are concatenated to form $\phi_c^\epsilon$ which is passed through a fully connected layer of size same as $\phi_c$ to produce the final feature for classification or re-identification.

$$
w_{\text{app}}^i = \exp \left( \frac{\phi_{c, i}^\text{app} \cdot \phi_{c, \text{max}}^\text{app}}{||\phi_{c, i}^\text{app}|| ||\phi_{c, \text{max}}^\text{app}||} \right) \quad (13)
$$

$$
w_{\text{flow}}^i = \exp \left( \frac{f_{c, i}^\text{flow} \cdot f_{c, \text{max}}^\text{flow}}{||f_{c, i}^\text{flow}|| ||f_{c, \text{max}}^\text{flow}||} \right) \quad (14)
$$

From Eqn. 13 and Eqn. 14 feature aggregation weights are calculated for both image features and flow features respectively. The features are aggregated as per weighted aggregation Eqn. 4 to form outputs $\phi_c$ and $f_c$ which are aggregated video features for image and flow respectively. These two features are concatenated to form $\phi_c^\epsilon$ which is passed through a fully connected layer of size same as $\phi_c$ to produce the final feature for classification or re-identification.

**4. Experimental Methodology**

In this section the experimental procedure is described to evaluate and compare the performance of proposed and reference feature aggregation methods. The implementation details describe the datasets, feature extraction models, parameters for flow feature estimation, optical flow estimation method, training and evaluation techniques.

**Datasets.** Experiment are performed on 2 challenging and widely-used datasets for video-based person reID. MARS [35] dataset is one of the largest datasets for video Person-ReID. It contains more than a 1000 identities and the capture conditions are close to practical applications.
The images are basically cropped bounding boxes. Another dataset commonly used in literature for evaluating video person ReID is Duke-MTMC [45, 30] dataset containing 702 identities and more than 2000 sequences for testing and training each.

**Deep Feature Extraction.** We follow the overall system architecture in [12] (Baseline) and [37] (COSAM). They achieved state-of-the-art results on several ReID datasets. They have used ResNet50 for feature extraction. We propose to use ResNet50 as our base network to learn features invariant to cluttered background by attending with saliency map obtained from optical flow estimations. The networks have been pre-trained on imagenet [10] 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 [12] and [37].

**Optical Flow Estimation.** To estimate optical flow maps for a given sequence LiteFlowNet [17] model has been chosen as they are computationally efficient compared to other deep models along with obtaining state of the art performance. We have used the official implementation from the authors to produce flow maps for both MARS and Duke-MTMC dataset. 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 $F^{t-1}$. Though the bounding boxes of these datasets are cropped and centered making it difficult to estimate accurate flow maps, we were able to produce visually acceptable flow maps with LiteFlowNet as shown the Figures earlier.

**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 extractor produces a $2048 \times 1$ 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.

**Settings.** The network feature extractors have been pre-trained on Imagenet dataset. We follow common data augmentation such as random flips and random crops 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 layers has been presented in the next section. Hence the training setting have mostly been kept similar to our baselines [12] and [37] for a fair comparison of results with baseline.

### 5. Results and Discussion

We start this section with some ablation study on contribution of different modules for the final result. We also include some empirical study on selecting the Deep Feature Extractor layer for for attention as well as sequence length selection. We then proceed to overall performance comparison with the baselines and finally a comparison with state-of-the-art methods. We do the ablation study on baseline ResNet50 architecture based system [12] with temporal pooling for feature aggregation on MARS dataset. We compare the overall performance on both MARS and Duke-MTMC dataset.

#### a) 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. Different layers in the Deep CNN have different abstraction

| Method                  | Feat Aggregation | mAP  | Rank 1 |
|-------------------------|------------------|------|--------|
| Baseline                | Pooling          | 75.8 | 83.1   |
| COSAM                   | Pooling          | 77.2 | 83.7   |
| Gated Attention (Simple Atten) | Pooling         | 77.4 | 84.6   |
| Mutual Attention (Mutual Atten) | Pooling        | 79.1 | 85.4   |
| Baseline                | RNN              | 75.7 | 82.9   |
| Baseline                | Temporal Attention | 76.7 | 83.3   |
| COSAM                   | Temporal Attention | 76.9 | 83.6   |
| Gated Attention (Simple Atten) | Weighted Addition | 77.8 | 84.8   |
| Mutual Attention (Mutual Atten) | Weighted Addition | 80.0 | 86.6   |

Table 1. Comparison of performance of baseline(ResNet50 + Temporal Pooling (TP)) with that of Baseline + Gated Attention + TP (Ours) at different layers of ResNet50 evaluated on MARS dataset.

Table 2. An ablation study of contribution of different module i.e Gated Attention(ours) and our Mutual Attention on the baselines. This study was done on MARS dataset.
Table 3. Empirical study of video sequence length vs performance on our method with Gated Attention, Mutual Attention, and baselines with no attention, average pooling and RNN on MARS dataset

| N Frame Seq | Baseline | RNN | Ours | Ours |
|-------------|----------|-----|------|------|
|             | Average Pooling | No Attention | Gated Attention | Mutual Attention |
|             | MAP | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP | Rank 1 | mAP | Rank 1 |
| 2           | 61.0 | 81.8 | 73.9 | 84.5 | 74.8 | 82.5 | 73.9 | 84.5 | 72.7 | 85.3 |
| 4           | 74.1 | 83.2 | 75.7 | 82.9 | 77.8 | 84.9 | 77.7 | 85.9 | 72.7 | 85.9 |
| 6           | 74.4 | 82.7 | -    | 77.6 | 84.5 | 79.2 | 85.8 | -    | 77.6 | 85.8 |
| 8           | 74.3 | 82.0 | 76.2 | 82.5 | 77.3 | 84.2 | 79.3 | 85.4 | -    | 77.6 | 85.8 |
| 16          | -     | -    | -    | -    | 72.5 | 82.9 | 80.0 | 86.6 | -    | -    | -    |

level of 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 [12] ReID system and evaluated on MARS dataset. From the Tab. [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 the earlier layers have different abstraction level of the salient features, the abstraction level increases in the deeper layers but in the last layer the spatial coherence is lower than the previous layers.

b) Contribution of Different Modules to the Baseline.

In this subsection we compare our Gated Fusion and Mutual Attention methods to the baseline [12] and that of state-of-the-art [37](ICCV-2019) since they follow similar comparison strategy as ours. In Tab. [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 can see that just Gated Attention on its own is [12] has improved the performances by a large margin compared to improvement by [37](COSAM-ICCV-2019) on MARS datasets. Our feature addition method introduced in Gated Attention method has further performed better than Average Pooling in [12]. 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.

c) 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 analyse the effect of sequence length on different feature aggregation methods such as Temporal Pooling. Flow guided weighted Addition applied on our method. Hence in Tab. [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 our 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.

d) Comparison with State-of-the-Art. We report the performance of our method ResNet50 + Gated attention with our weighted addition feature aggregation method and also our system with Mutual Attention and weighted feature aggregation method on MARS and Duke-MTMC datasets in the Tab. [4] We also compare our results with some of the related state-of-the-art methods in the table. As mentioned earlier we have selected [12] 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 performed most of the state-of-the-art methods including some of the best existing methods. We have also shown the advantage of our method compared to other optical flow base methods [47, 44, 6]. Although [22] 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 optimal flow and image stream can attend to the salient regions of each other.

6. Conclusion

In this work we present two novel frameworks for flow guided attention and temporal feature aggregation for Person-ReID. The sole purpose of the work has been to focus on common saliency in moving objects to reduce background clutter, encode motion information to learn motion patterns of person that enable to have the advantages of having longer frame 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 best recent state-of-the-art in Rank1 metric in addition to outperforming other methods on both mAP and Rank1 metrics, evaluated on MARS as well as Duke-MTMC dataset.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In CVPR, 2015.

[2] Amran Bhuiyan, Yang Liu, Parthipan Siva, Mehrsan Javan, Ismail Ben Ayed, and Eric Granger. Pose guided gated fusion for person re-identification. In The IEEE Winter Conference on Applications of Computer Vision, pages 2675–2684, 2020.

[3] Amran Bhuiyan, Alessandro Perina, and Vittorio Murino. Person re-identification by discriminatively selecting parts and features. In ECCV, 2014.

[4] Amran Bhuiyan, Alessandro Perina, and Vittorio Murino. Exploiting multiple detections for person re-identification. Journal of Imaging, 4(2):28, 2018.

[5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduardo Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In NIPS, 1994.

[6] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1169–1178, June 2018.

[7] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In CVPR, 2017.

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In CVPR, 2016.

[9] D. Chung, K. Tahboub, and E. J. Delp. A two stream siamese convolutional neural network for person re-identification. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 1992–2000, Oct 2017.

[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.

[11] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In CVPR, 2010.

[12] Jiayang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. arXiv preprint arXiv:1805.02104, 2018.

[13] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. In ICCV, 2019.

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

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[16] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7132–7141, June 2018.

[17] Tak-Wai Hui, Xiaou Tang, and Chen Change Loy. Lite-flownet: A lightweight convolutional neural network for optical flow estimation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8981–8989, 2018.

[18] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712, 2018.

[19] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In CVPR, 2017.

[20] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multiscale 3d convolution network for video based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 8618–8625, 2019.

[21] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In CVPR, 2015.

[22] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In British Machine Vision Conference, 2019.

[23] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing, 26(7):3492–3506, 2017.

[24] Yu Liu, Yan Junjie, and Wanli Ouyang. Quality aware network for set to set recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[25] Y. Li, Meng Ji, Zhipeng Lai, and Yufeng Liu. Improved R-CNN for person re-identification. In The IEEE Winter Conference on Applications of Computer Vision, 2017.

[26] N. McLaughlin, J. M. d. Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1325–1334, June 2016.

[27] N. McLaughlin, J. M. d. Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1325–1334, June 2016.

[28] Rameshwar Pandey, Amran Bhuiyan, Vittorio Murino, and Amit K Roy-Chowdhury. Unsupervised adaptive re-identification in open world dynamic camera networks. In CVPR, 2017.

[29] Xueling Qian, Yanwei Fu, Tao Xiang, Wenzhan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In ECCV, 2018.
[29] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. *ICCV*, 2019.

[30] 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 workshop on Benchmarking Multi-Target Tracking*, 2016.

[31] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *ICCV*, 2018.

[32] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5363–5372, June 2018.

[33] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[34] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, June 2018.

[35] Springer. MARS: A Video Benchmark for Large-Scale Person Re-identification, 2016.

[36] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[37] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[38] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. *ECCV*, 2018.

[39] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *European Conference on Computer Vision (ECCV)*, 2018.

[40] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.

[41] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019.

[42] Pavan Turaga, Rama Chellappa, and Ashok Veeraraghavan. Advances in video-based human activity analysis: Challenges and approaches. In Marvin V. Zelkowitz, editor, *Advances in Computers*, volume 80 of *Advances in Computers*, pages 237 – 290. Elsevier, 2010.

[43] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[44] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, 2020.

[45] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[46] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[47] Ruimao Zhang, Jingyu Li, Hongbin Sun, Yuying Ge, Ping Luo, Xiaogang Wang, and Liang Lin. Scan: Self- and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*, 28(10):4870–4882, 2019.

[48] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[49] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *CVPR*, 2017.

[50] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.

[51] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[52] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6776–6785, July 2017.