Image and Vision Computing 123 (2022) 104474

Contents lists available at ScienceDirect

Image and Vision Computing

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

STCA: Utilizing a spatio-temporal cross-attention network for enhancing video person re-identification

Amran Bhuiyan [a,b,*], Jimmy Xiangji Huang [a,*]

[a] Information Retrieval and Knowledge Management Research Lab, School of Information Technology, York University, Toronto, Canada
[b] Deptartment of CSTE, Noakhali Science and Technology University, Noakhali, Bangladesh

ARTICLE INFO

Article history:
Received 1 February 2022
Received in revised form 13 April 2022
Accepted 9 May 2022
Available online 17 May 2022

Keywords:
Re-identification
Deep learning
3D-CNNs
Cross attention

ABSTRACT

Video-based re-identification (ReID) is a crucial task in computer vision that draws increasing attention due to advances in deep learning (DL) and modern computational devices. Despite recent success with CNN architectures, single models (e.g., 2D-CNNs or 3D-CNNs) alone failed to leverage temporal information with spatial cues. This is due to uncontrolled surveillance scenarios and variable poses leading to inevitable misalignment of ROIs across the tracklets, which is accompanied by occlusion and motion blur. In this context, designing temporal and spatial cues for two different models and their combinations can be beneficial, considering the global of a video-tracklet. 3D-CNNs allow encoding of temporal information while 2D-CNNs extract spatial or appearance information. In this paper, we propose a Spatio-Temporal Cross Attention (STCA) network to utilize both 2D-CNNs and 3D-CNNs that calculate the cross attention mapping both from the layer of 3D-CNNs and 2D-CNNs along a person's trajectory to gate the following layers of 2D-CNNs; and highlight relevant appearance features for the person ReID. Given an input tracklet, the proposed cross attention (CA) is able to capture the salient regions that propagate throughout the tracklet to obtain the global view. This provides a spatio-temporal attention approach that can be dynamically aggregated with spatial features of 2D-CNNs to perform finer-grained recognition. Additionally, we exploit the advantage of utilizing cosine similarity while triplet sampling as well as for calculating the final recognition score. Experimental analyses on three challenging benchmark datasets indicate that integrating spatio-temporal cross attention into the state-of-the-art video ReID backbone CNN architecture allows for improving their recognition accuracy.

1. Introduction and motivation

Person Re-Identification (ReID) is the task of associating people across a set of non-overlapping camera viewpoints. It has drawn much attention from the computer vision community due to its vast application area, ranging from video surveillance to sports analytics. Much progress has been made in developing successful deep learning (DL) [1–11], yet it consists of a challenging task due to the nonrigid structure of the human body, different viewpoints, and poses with which a pedestrian can be observed, occlusion, and misalignment issues.

Existing ReID techniques can be categorized into two categories: image-based ReID [1–4,6–8,12–14] and video-based ReID [15–24]. In the former case, query images (still) of different individuals are matched with the gallery images (still) without considering the temporal information. In video-based ReID settings, a person's small input video sequence (i.e., tracklet) is matched against the gallery of small video sequence (i.e., tracklet) representations that consider the temporal aspect of different time steps. In image-based ReID setting, state-of-the-art approaches [1,6–8,12–14] mostly rely on appearance-based features that capture only the color or texture of the clothes. Relying only on appearance features might harm the recognition accuracy due to misaligned bounding boxes or occlusions. In contrast, by taking video sequence as input, the video-based ReID setting has the advantage of having additional information, which allows the exploitation of spatio-temporal information for precise, comprehensive and discriminative feature embedding from a global point of view.

Most of the video-based ReID approaches model the temporal relation by employing 3D-CNNs [18,20,22], temporal pooling [15], temporal attention [15], spatio-temporal attention mechanism [16,21,23–26]. Most of these approaches treat the video-based ReID problem as other video-based computer vision tasks (e.g., action or activity recognition) by ignoring the fact that video clips in ReID consist of a sequence of consecutive bounding boxes (i.e., region of interests (ROIs)) produced by a person detector, not the original video frames of the whole captured scene. The inefficiency of the person detection algorithms lead to

* Corresponding authors at: Information Retrieval and Knowledge Management Research Lab, School of Information Technology, York University, Toronto, Canada.
E-mail addresses: amran@yorku.ca (A. Bhuiyan), jhuang@yorku.ca (J.X. Huang).

https://doi.org/10.1016/j.imavis.2022.104474
0262-8856/

Fig. 1. Examples of the misalignment challenge characterizing person re-identification over two camera viewpoints from the MARS dataset. Each row represents the bounding boxes from the same individual. The first four columns of each row are from the same tracklet captured with the first camera '**X**' while the remaining four columns are from another tracklet captured with the second camera'**Y**'.

extracting smaller or bigger bounding boxes compared to the ground truth as depicted in Fig. 1, known as *misalignment* issues. The sources of misalignment generate both from spatial and temporal issues [22]. It is apparent from Fig. 1 and the statement from the appearance preserving approach [22] that inherent challenges such as pose or viewpoint variations may lead to spatial misalignment issues. There are cases when a probe image with a certain alignment is matched with a gallery of different images with different alignments, leading to a low recognition score. In temporal misalignment issues, the same spatial locations in consecutive frames from different body parts and the same body parts in adjacent frames may be scaled to different sizes. In such cases, using only simple temporal pooling as a feature aggregation technique may lead to poor recognition accuracy. Temporal information extracted using only 3D-CNNs in such scenarios may destroy the appearance representations.

There are a number of state-of-the-art ReID approaches that deal with misalignment issues by deploying either attention mechanisms [21,25,26] or preserving the appearance of 3D-CNNs [22]. Relying only on single 2D-CNNs, most of the attention guided video ReID approaches [21,25,26] model the attention in a 2D fashion where the output of 2D-CNNs layers are transformed through an external module (i.e., pose [7] and segmentation network [25]) to yield the final attention map. Yet, they failed to capture common or global temporal attention that is uniform throughout the tracklet. There have been some attempts to leverage spatio-temporal attention [16,24,27] for video-based ReID that model the spatial and temporal attention separately and sequentially using complex dedicated structures. However, they ignore the effect of learning temporal attention that provides a global view of a tracklet while calculating spatial attention. The extracted feature embedding in a such scenario suffers from the issue of accurately positioning the feature saliency considering the degree of redundancy present in the entire tracklet. Thus, a dynamic model is expected that can jointly learn the salient levels of each spatio-temporal feature from the global point of view.

Within this context, we propose a STCA framework that utilizes both 2D-CNNs and 3D-CNNs for generating cross guided global spatio-temporal attention maps from the global point of view to deal with spatial and temporal misalignment issues, respectively. In this context, an important research query is how and where to combine both feature representations. To answer this query, we propose to use the cross attention approach by processing the outputs from the same convolutional layer of 3D-CNNs and 2D-CNNs to gate the output of 2D-CNN backbone layers. The proposed STCA approach comprises two parallel streams, a backbone 2D-CNNs to learn spatial appearance information and a 3D-CNNs to model the temporal information for ReID. These streams are combined using the cross attention network to learn spatio-temporal attention into the metric learning process. The proposed cross attention network jointly learns the feature embedding that relies on a simple cross attention mechanism that allows for multiplicative interaction between the 2D input features and cross attention map, providing a dynamic selection of consistent filters throughout an input tracklet. Dynamic selection of the discriminant filters initiated through cross attention mechanisms avails the 2D backbone to perform fine-grained recognition which in turn helps to address the misalignment issues. Most of the attention networks for ReID approaches [7,16] apply the attention map into the later parts of the backbone network that raises an important research question about the most relevant layers for application of the attention map. There are a few approaches in ReID [14,25] to exploit the advantage of fusing mid-level or contextual feature (i.e. pose [14] and segmentation mask [25]) into the mid-level layers. Unlike these approaches [14,25], we propose using cross attention map to gate the mid-level layer of the 2D-CNN backbone layer that provides back-propagated gradients corresponding to the amplified local similarities across the tracklets.

The DL frameworks for most of the state-of-the-art video ReID approaches [15,16,22,24] are optimized using the combination of triplet loss and classification loss. It is of utmost importance mining proper triplet samples to obtain faster convergence and better recognition accuracy. State-of-the-art video ReID [6,15,16,22] approaches relied on Euclidean distance for sampling batch-hard positive and negative mining that failed to obtain faster convergence and better recognition accuracy. Unlike those approaches [6,15], we rely on cosine distance for mining semi-hard positive and negative samples in a batch.

The contribution of this paper is three-fold:

(1) We propose a novel DL framework (STCA) that includes both 2D-CNNs and 3D-CNNs for generating cross guided global spatio-temporal attention maps for gating 2D-CNN backbones in video-based ReID. Given an input tracklet, it allows leveraging common salient features consistent throughout space and time, thereby mitigating the spatio-temporal misalignment issue.
(2) We exploit the advantage of using cosine distance while mining positive and negative samples for semi-hard triplet sampling. Given an anchor in a batch, finding the furthest sample as hard positive and the closest sample as hard negative using Cosine distance for triplet sampling leads to faster optimization and better recognition accuracy.
(3) We conduct an extensive experimental analysis to validate the proposed approach on three video-based ReID datasets, allowing us to conclude that our cross attention between 2D- and 3D- into 2D-CNN backbone can outperform most of the state-of-the-art methods for video ReID.
The code to reproduce the results of this paper is published on GitHub.[1]

2. Related work

There are a large number of studies on the topics of ReID [1,8,13,14,28–33] both in image-based and video-based setting. Here we mainly review the work about video-based ReID, which is the most related to our research presented in this paper.

2.1. Video-based ReID

Since the advancement of modern computational infrastructure, video-based ReID attracted much attention as modeling the temporal information allows the model to deal with the issues such as occlusion, background noise and misalignment. A common practice for video ReID is a two steps process: extracting feature embedding from each frame and then aggregating the embeddings of all the frames in a tracklet using temporal average/max-pooling [7,15,34]. There have been a few attempts [35,36] to characterize the temporal variations of each time-step. In [35], a variant of RNN is proposed, which is followed by temporal pooling to produce the final feature representation of a given tracklet. Following the same pipeline, a LSTM network is proposed in [36] to aggregate the features that obtain a sequence level feature representation. Similarly, a RNN based approach is proposed in [37] to combine a motion context with the corresponding appearance representations. Recent success in other video-related tasks (such as action recognition) with 3D-CNNs has prompted the ReID community to apply it for video-based ReID [18,20,22]. In this context, the spatio-temporal representation learned with a 3D-CNN has been concatenated with appearance representation with 2D-CNN [20]. In a similar fashion, 3D-CNNs [18] have been combined with non-local blocks for spatio-temporal attention, allowing to model the long-range dependencies [38]. More recently, a 3D-CNN based robust appearance preserving approach [22] is proposed to deal with the misalignment issues at the pixel level. However, all of the approaches mentioned above process the features with the same importance even using a single CNN though the features for different spatial and temporal positions have different levels of significance in video-based person ReID.

2.2. Attention based video ReID

The task of attention-based approaches is to highlight the local regions to extract more discriminative features. Similar to other computer vision applications, it has achieved great success in video ReID as well. State-of-the-art attention base video ReID [16,21,27,39] can be categorized into two approaches. One category of attention mechanism relies on the relative distance of consecutive frames [19,40,41]. In this category [19,41], authors have proposed using matching a given probe sequence against all gallery sequences based on a distance measure with temporal attention weights. Similarly, inter-and intra-tracklet level comparisons are performed in [40] to design dual attention mechanisms. However, such distance-based attention mechanisms become impractical when the number of gallery instances grows since they calculate each gallery's different features w.r.t. a given probe instance. On the other hand, in the second category of attention mechanism, one sequence doesn't depend on other sequences to calculate the attention map [16,21,24,42–48]. In this case, for each frame, there is a learned quality score from the quality-aware network [49] before aggregating them into the final feature representations of a given sequence. Similarly, Zhou et al. [50] learn temporal attention to exploit the feature importance of each frame and update it using a RNN. With the attribute-driven weighing mechanism, the authors [39] leverage attention weighing that is driven by a confidence score on the attribute recognition. Nonetheless, all these approaches rely entirely on temporal attention by ignoring the importance of using spatial attention. There have been a few attempts to learn spatio-temporal attention using the same output of the 2D-CNN or the 3D-CNN backbone where attention is learned as prior knowledge, although that limits their use in optimization problems [16,21,24,42–48]. Using a single 2D-CNN back to learn both the spatial and temporal attention separately or sequentially [16,21,24,42–44,48] hinders the process of exploiting mutual benefits between spatial and temporal features during optimization. Recently, there has been an attempt to use the optical flow that is fused with appearance cues in a mutual fashion [23]. However, using optical flow as a parallel stream with the same importance as RGB-stream [23] is less effective as optical flow only illustrates coarse semantic features of moving objects, and thus not compatible while treating equally with RGB-like appearance feature for recognition.

Different from all the above approaches, we proposed a STCA framework that generates cross attention by taking the input from the common layer of 3D-CNN and 2D-CNN to gate the feature representations into the following layers of 2D-CNN. Focusing on the state-of-the-art DL framework, the proposed cross attention learns the global spatio-temporal attention for a given tracklet, enabling the 2D backbone to pay more attention to the common salient feature representations consistent throughout the tracklet.

3. Our proposed approach

This section describes the proposed STCA framework that generates spatio-temporal cross attention maps to guide the output of 2D-CNN backbone layers for video ReID. Fig. 2 illustrates the layout of the proposed model. It consists of three blocks: a 2D-CNN backbone, 3D-CNN temporal backbone, and a cross attention module. All these three components are jointly trained end-to-end in a single step. Given an input tracklet, the features produced by the 2D-CNN and 3D-CNN layers are used by the attention network to guide the following layer of the 2D-CNN backbone so that it could dynamically select the most discriminant convolutional filters from the backbone to ensure fine-grained recognition.

3.1. Backbone 2D-CNN architecture

As appearance based feature representation plays a key role in the recognition process, choosing a 2D-CNN is vital. There are a number of 2D-CNN architectures such as: VGG [51], ResNet [5], ResNet-IBN [52], SE-ResNet [53], and Inception [54]. Among them, ResNet [5] and its variants are commonly used for the application of ReID. The idea of using skip connection in ResNet [5] allows propagating the original input data as deep as possible through the network while addressing the

[1] Code available at: https://github.com/mdamranhossenbhuiyan/STCA.

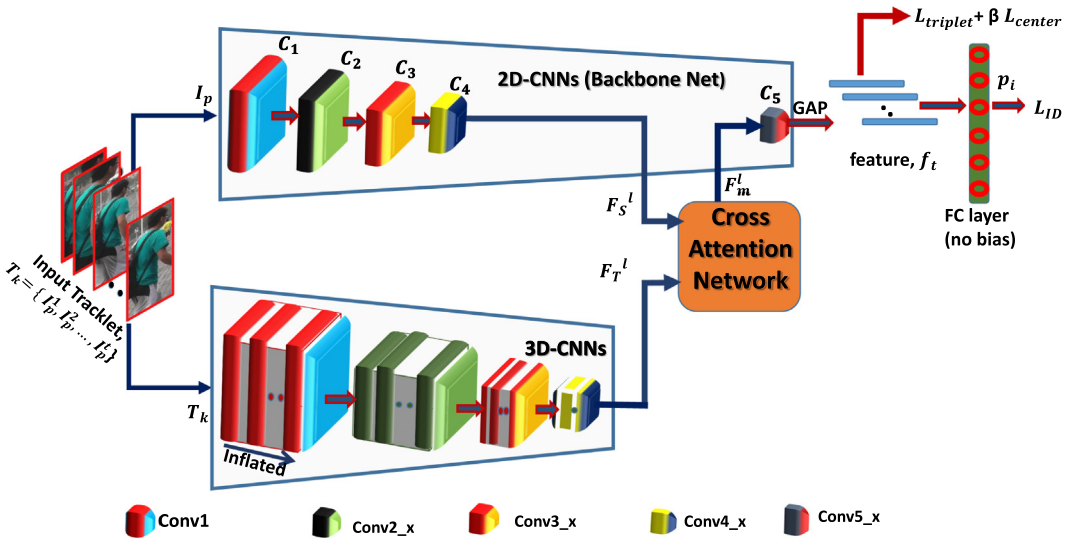

Fig. 2. Proposed STCA network architecture for video person ReID. The 3D-CNN learns the spatio-temporal features while 2D-CNN learns appearance features. The cross attention network uses these features to dynamically select the relevant 2D-CNN appearance features. It allows the 2D-CNN backbone to learn common salient appearance cues for the embedding space across a tracklet.

issues of vanishing gradient in VGG [51]. The combination of deeper architecture due to skip connection as well as the ensemble nature of multiple staked convolutional blocks in ResNet [5] assist the network to learn robust feature embedding for higher recognition accuracy. All of ResNet variants comprise multiple convolutional blocks, for example ResNet [5] architecture is divided into one convolutional block named *conv1- C1*, followed by four residual blocks named *conv2_x - C2, conv3_x - C3, conv4_x - C4* and *conv5_x - C5*, respectively, as shown in the backbone network structure of Fig. 2. Fragmented architectures of this kind allow us to analyze layer-wise feature representations.

Given an input tracklet (set of bounding boxes extracted from a tracklet, T_k) represented by $\mathbf{I_p^1}$,$\mathbf{I_p^2}$, ...,$\mathbf{I_p^t}$ where p indicates the identity of the video sequence of length t, the backbone 2D-CNN, Φ computes the appearance features, $\mathbf{F}_A^l$, which is the output of the l-th layer of the backbone network for the batch of input images, $\mathbf{I_p}$:

$$\mathbf{F}_A^l = \Phi^l(\mathbf{I_p}) \tag{1}$$

where Φ^l is the appearance feature extractor until l-th layer define as $\Phi^l : \mathbf{I_p} \rightarrow \mathbf{F}_A^l$, $\mathbf{I_p} \in \mathbb{R}^{b\times 3\times h\times w}$, $\mathbf{F}_A^l \in \mathbb{R}^{b\times c_l\times h'\times w'}$, and $b = p \times t$ being the batch size, h, and w being the height and width of an input image and h', w', c_l being the height, weight and channel number of attention map size of the l-th layer. Reshaping the batch dimensions b into P and t will provide the feature representations with temporal dimension, $\mathbf{F}_A^l \in \mathbb{R}^{p\times t\times c_l\times h'\times w'}$.

3.2. 3D-CNN architecture

A cross attention network typically receives two signals: one from 2D-CNN and another from 3D-CNNs. 2D-CNN typically extracts appearance-based spatial features while 3D-CNN encodes temporal information. Therefore, choosing an efficient 3D-CNN is also important for our proposed approach. Several commonly used 3D-CNN architectures are C3D [55], I3D [56] and 3D-ResNet [57] that can be employed to model the temporal variations. One group of 3D-CNN designs the convolutional network [55,57] with a filter kernel of size $3 \times 3 \times 3$. Another group of 3D-CNN turns a pre-trained 2D Network architecture to 3D-CNN by inflating all the filters and pooling kernels with an additional temporal dimension. One of the above-mentioned 3D-CNN architectures can be used to get the temporal features and thus can be used as the input of the cross attention network.

In our approach, the input of 3D-CNN is the same batch of the backbone architecture with the same size but different shape as it takes temporal or tracklet, T_k information into account in the additional dimension. The 3D-CNN, Ψ computes the temporal variations, $\mathbf{P}_t^l$, which is the output of the l-th layer of the backbone network for the batch of input tracklets, $\mathbf{T}_k$:

$$\mathbf{F}_T^l = \Psi^l(\mathbf{T}_k) \tag{2}$$

where, Ψ^l is the temporal feature extractor until l-th layer define as $\Psi^l : \mathbf{T}_k \rightarrow \mathbf{F}_T^l$, $\mathbf{T}_k \in \mathbb{R}^{p\times t\times 3\times h\times w}$, $\mathbf{F}_T^l \in \mathbb{R}^{p\times c_l\times h'\times w'}$, where $b = p \times t$ is the batch size, h, and w is the height and width of an input image and h', w', c_l being the height, weight and channel number of attention map size of the l-th layer.

3.3. Cross attention network

The Cross attention network utilizes both the 2D-CNN and 3D-CNN stream that attends each other to obtain the semantic relevance that enables the 2D-CNN backbone to learn the common salient features consistent throughout the sequence of a given tracklet. The proposed network is inspired by the cross-attention approach [58] that has been adapted for the application of video ReID.

As shown in Fig. 3, the cross attention network takes two inputs: appearance features, $\mathbf{F}_S^l$ from the l-th layer of backbone architecture and the temporal features, $\mathbf{F}_T^l$ from 3D-CNN. This module generates the cross attention map A_S for $\mathbf{F}_S^l$, which is then used to weight the feature map into the following layer of the 2D-CNN to achieve a more discriminative feature representation $\mathbf{F}_m$. For simplicity, we ignore subscript and superscript and denote the 2D-CNN and 3D-CNN feature maps as $\mathbf{F}_S$ and $\mathbf{F}_T$.

Firstly, we devise a correlation layer to calculate a correlation between $\mathbf{F}_S$ and $\mathbf{F}_T$ to guide the generation of cross attention maps. To do this, we first reshape the $\mathbf{F}_S$ and $\mathbf{F}_T$ to $\mathbb{R}^{b\times c_l\times k}$, where $k = h' \times w'$ is the number of spatial and temporal locations for $\mathbf{F}_S$ and $\mathbf{F}_T$, respectively. Then, we calculate the cosine distance between each point of the spatial location of $\mathbf{F}_S$ with all the points of the temporal locations of $\mathbf{F}_T$ that gives us the spatial correlation map, $R_s \in \mathbb{R}^{k\times k}$. R_s represents the local correlations between the feature maps of 2D-CNN and 3D-CNN. Then, a convolutional operation is performed on R_s with a $k \times 1$ kernel, $w \in \mathbb{R}^{k\times 1}$ to fuse each local correlation vector. w plays a vital

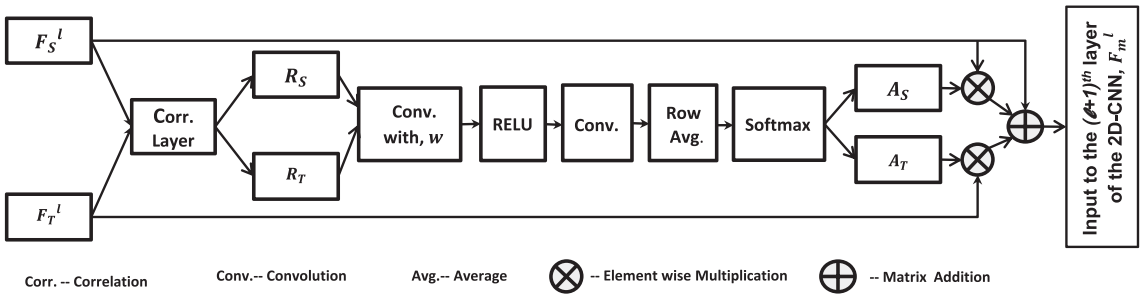

Fig. 3. A schematic illustration of the cross attention network.

role as it aggregates the correlation between $\mathbf{F}_S$ and $\mathbf{F}_T$. To learn it adaptively, we apply RELU to introduce non-linearity in the learning that allows flexible transformation, followed by a convolutional operation and row-wise averaging. The resultant output is passed through a softmax function to normalize it to obtain the final attention map $\mathbf{A}_S$.

In a similar fashion, we can obtain $\mathbf{A}_T$. A residual attention mechanism is applied both for $\mathbf{A}_S$ and $\mathbf{A}_T$, where the initial feature maps $\mathbf{F}_S$ and $\mathbf{F}_T$ are element-wisely weighted by $1 + \mathbf{A}_s$ and $1 + \mathbf{A}_T$ respectively to obtain cross attended feature, $\mathbf{F}_m^l$ to gate the following 2D-CNN layers. In addition to the learned attention, the residual mechanism enables the network to focus more on semantically important regions and thus improves its ability for discriminative representation. This in turn helps to address the issue of misalignment.

3.4. Network loss with cosine distance

Most of the state-of-the-art video ReID approaches [15,16,20,22] optimize their proposed framework using the combination of triplet and identity loss. Unlike those, our proposed STCA framework is trained with the techniques adopted in the bag-of-trick(BOT) [8] approach that combines triplet loss, center loss, and identity (ID) loss with label smoothing. For a given mini-batch of input tracklets, the total loss can be written as:

$$\mathcal{L} = \mathcal{L}_{triplet} + \mathcal{L}_{ID} + \beta\mathcal{L}_{center} \tag{3}$$

where, $\mathcal{L}_{triplet}$, $\mathcal{L}_{ID}$, and $\mathcal{L}_{center}$ denote the triplet loss, ID loss, and center loss, respectively. β is the weight of the center loss that balances it with other losses. Following BOT [8], we set β is set to 0.0005.

As a triplet loss, we utilize widely used hard triplet mining in a mini-batch [6], the task of which is to find hard positive and negative to form a triplet during the training process. As pointed-out by Hermans et al. [6], in a mini-batch, optimizing the network for similarly-looking individuals but different identities (hard negative) or different-looking individuals but the same identity (hard positives) learns to distinguish better than randomly chosen triplet sampling from the whole dataset. For each sample in a mini-batch, it selects the furthest positive sample as the hard positive and the closest negative sample as hard negative samples within the batch using Euclidean distance when forming the triplets for computing the loss. Different from their approach [6], the proposed STCA approach utilizes the *cosine distance, d,* while forming triplet which can be formulated as:

$$\mathcal{L}_{triplet} = \frac{1}{N_s}\sum_{a=1}^{N_s}\left[m + \max_{[y_p=y_a]} d(\mathbf{f}_{ta}, \mathbf{f}_{tp}) - \min_{[y_p\neq y_a]} d(\mathbf{f}_{ta}, \mathbf{f}_{tn})\right]_+ \tag{4}$$

where, $[.]_+ = \max(.,0)$, m denotes a margin, N_s is the set of all hard triplets considering *cosine distance, d,* within a batch. $\mathbf{f}_{ta}$, $\mathbf{f}_{tp}$ and $\mathbf{f}_{tn}$ denote the feature representations of the anchor, positive and negative samples with their labels y_a, y_p and y_n, respectively.

To characterize the intra-class variations, minimizing the center loss [59] has proven to be effective that penalizing the distances between the deep features and their corresponding class center. In this way, it helps to regularize the triplet loss so that it doesn't overfit the training ID The center loss can be formulated as:

$$\mathcal{L}_{center} = \frac{1}{2}\sum_{j=1}^{N}\|\mathbf{f}_{t_j} - \mathbf{c}_{y_j}\|_2^2 \tag{5}$$

where, y_j is the label of the j-th image in a mini-batch. c_{yj} denotes the y_j th class center of extracted features. N is the number of samples in the batch.

To prevent the ReID model from overfitting training IDs, integration of label smoothing (LS) [60] into ID loss is a widely used technique. Given an image of an individual with a true ID label, y and p_i as ID prediction logits of class i, ID loss can be written as:

$$\mathcal{L}_{ID} = \sum_{i=1}^{N} -q_i \log(p_i) \tag{6}$$

where LS construction of q_i is:

$$q_i = \begin{cases} 1 - \dfrac{N-1}{N}\varepsilon, if \;\; i = y \\ \dfrac{\varepsilon}{N}, otherwise \end{cases} \tag{7}$$

where, ε is a constant to force the model to be less confident on the training set. Following the BOT [8] approach, ε is set to be 0.1. For small datasets with fewer IDs, LS has proven to be effective in improving the model's performance.

4. Experimental methodology and result analysis

4.1. Datasets

The proposed STCA framework is evaluated on 3 challenging benchmark datasets: MARS [61], DukeMTMC-VideoReID [62], iLIDS-VID [63]. **MARS** [61] is one of the largest sequence-based person ReID datasets that consists of 1261 identities and 20,478 video sequences, with multiple frames per person captured across 6 non-overlapping camera views. Following the state-of-the-art techniques [15,16,25,61], 625 identities are used for training, and the remaining are used for testing. **DukeMTMC-VideoReID** [62] contains 4832 tracklets from 1812 identities captured across 8 synchronized cameras with a standard train/test protocol where 702 identities are used for training, 702 for testing and remaining 408 identities are distractors. **iLIDS-VID** [63] is a smaller dataset with 600 tracklets of 300 identities from two non-overlapping camera views. We followed the test protocol of [15,22,25,63] and 10 random probe-gallery splits are used to perform experiments.

Table 1
Accuracy of the proposed STCA and state-of-the-art methods on MARS, Duke: DukeMTMC-Video, and iLIDS: iLIDS-VID datasets.

Method	MARS		Duke		iLIDS
	mAP	rank-1	mAP	rank-1	rank-1
CNN + XQDA [61]	47.6	65.3	–	–	–
TriNet [6]	67.7	79.8	–	–	–
Part-Alignment [7]	72.2	83.0	78.3	83.6	–
Mask-Guided [9]	71.1	77.1	–	–	–
Snippet [19]	71.1	82.1	–	–	85.4
RevisitTemPool (Baseline) [15]	75.8	83.1	–	–	–
M3D [20]	74.1	84.4	–	–	74.0
coSAM (SE-ResNet50) [25]	79.9	84.9	94.1	95.4	–
STA [16]	80.8	86.3	94.9	96.2	–
AttDriven [39]	78.2	87.0	–	–	86.3
GLTR [70]	78.5	87.3	93.7	96.3	86.0
Mutula Attention [23]	80.9	87.3	94.8	96.7	–
VRSTC [17]	82.3	88.5	93.5	95.0	83.4
NVAN [66]	82.8	90.0	94.9	96.3	–
AP3D [22]	85.1	90.1	95.6	96.3	86.7
MG-RAFA [21]	85.9	88.8	–	–	–
I2GNN [67]	86.3	89.1	–	–	–
STCAN [24]	84.5	88.9	–	–	–
BiCnet-TKS [68]	86.0	90.2	96.1	96.3	–
STMN [69]	84.5	**90.5**	95.2	**97.0**	–
STCA (Ours)	**87.0**	90.3	**96.2**	96.6	**88.3**

4.2. Network architecture

Our proposed approach is evaluated by resizing all the images to 384×128 and 256×128. Although the proposed method can integrate a wide range of feature extractors, we choose the ResNet50-IBN [52] architectures as the backbone 2D-CNN due to their popularity in re-identification. However, we considered two modifications to adapt them for aligned feature representation. The quality of aligned features heavily relies on a sufficient resolution for the final activation. Thus we changed the stride of the last convolutional layer of all the backbone 2D-CNNs from 2 to 1. Then, the final layer is removed to provide a 2048-dimension feature representation. To remain consistent with the feature resolutions, we choose ResNet3D-50 [57] architecture as the 3D feature extractor. The ResNet50-IBN backbone and ResNet3D-50 networks are initially pretrained on ImageNet [64] and Kinectics [65] datasets, respectively. The cross-attention network is training from scratch which is randomly initialized from Gaussian distribution. The batch size is 32 that includes 8 identities with 4 video sequences. As for the optimizer, we use the Adam optimizer with weight decay 0.0005 was adopted to update the parameters for the optimizer. The learning rate is initialized to 3.5×10^{-4} and multiplied by 0.1 after every 10 epochs.

4.3. Performance measures

The proposed STCA model is evaluated for its ability to provide discriminant feature embeddings for accurate pairwise matching (with a Siamese network). Features extracted from query and gallery images for all the tracklets are averaged to get a single feature vector for each tracklet and then compared through pairwise matching. The similarity between each pair of feature embeddings is measured using cosine distance. For each query image, all gallery images are ranked according to the similarity between their embeddings in cosine space, and the most similar gallery image label is returned. All the experiments are implemented in PyTorch. Following the common trend of evaluation [15,16,22], we measure the rank-1 accuracy of cumulative matching characteristics (CMC), and the mean average precision (mAP) to evaluate our proposed and baseline methods. The CMC represents the expectation of finding a correct match in the top n ranks.

4.4. Comparison with state-of-the-art methods

This experiment aims to compare our proposed STCA approach with state-of-the-art methods. We compare with relevant models for video ReID, including alignment-based models: AP3D [22], Mask Guide [9], coSAM [25] and Part-Alignment [7]; attention guided models: STA [16], RevisitTemPool [15], VRTC [17], NVAN [66], MG-RAFA [21], AttDriven [39], I2GNN [67], STCAN [24], BiCnet-TKS [68], STMN [67] and others on MARS, Duke-MTMC and iLIDS-VID datasets. Table 1 reports the comparison of our proposed STCA approach with most of the state-of-the-art approaches (see Table 1). The following observations can be made: i) the proposed STCA framework for video ReID consistently outperforms all compared state-of-the-art methods on all datasets in mAP measures, which is a more comprehensive measurement to explain the overall effectiveness of an approach while the dataset has multiple ground-truths for each query. Although the STMN [69] approach obtains state-of-the-art performance in rank-1 measurement, the significant performance drop (2.5% and 1.0% on MARS and DukeMTMC-Video, respectively) in mAP measurement compared to our proposed STCA approach. Similar observations can be made on BiCnet-TKS [68], STCAN [24], I2GNN [67], and AP3D [22] while drawing a comparison with our STCA approach. ii) Among all the alternatives, STMN [69] obtains a strong performance on rank-1 which shows a marginal performance gap (0.2% and 0.4% on MARS and DukeMTMC-Video dataset, respectively) with our STCA approach. This may be because STMN adopts a complex LSTM based memory module for focusing on discriminative frames and achieving high performance on rank-1. iii) On the iLIDS-VID dataset, we outperformed all the considered state-of-the-art in rank-1 accuracy. This might be due to the availability of numerous misalignment data due to the severe viewpoint variations across camera views, cluttered background and random occlusions. iv) The performance of state-of-the-art methods varies significantly depending on their backbone networks and the different mechanisms they rely on. Compared to the baseline RevisitTempool [15], on which we apply our STCA approach, we are outperforming by 11.2% in mAP and 7.2% in rank-1 measures. Similar observation can be made for CNN + XQDA [61], TriNet [6], Mask Guide [9], coSAM [25], Part-Alignment [7], STA [16], RevisitTemPool [15], VRTC [17], NVAN [66], MG-RAFA [21], AttDriven [39], compare to which our STCA approach significantly outperformed. Nevertheless, our proposed approach consistently outperforms the other state-of-

Table 2
Influence of varying resolutions on MARS dataset.

Method	256×128		384×128		
	mAP	rank-1	mAP	rank-1	epoch
3D-CNN [17]	70.5	78.1	73.5	81.3	≈ 800
3D-CNN with Cosine	75.2	83.6	77.2	85.9	≈ 500
2D-CNN [17]	75.8	83.1	77.4	84.5	≈ 800
2D-CNN with Cosine	81.3	87.1	82.2	87.9	≈ 500
Concatenate(2D-CNN,3D-CNN) with Cosine	81.9	86.9	83.1	86.6	≈ 500
STCA with Cosine	**84.9**	**89.1**	**87.0**	**90.3**	≈ 450

Table 3
Accuracy of the proposed approach using ResNet50-IBN with ensembling versus the sequence length on the MARS dataset.

Frame Length	Accuracy			
	mAP	rank-1	rank-5	rank-10
2	86.5	89.4	97.4	97.8
4	**87.1**	90.1	97.6	**98.3**
6	87.0	89.9	**97.8**	98.3
8	87.0	**90.3**	97.6	98.2
10	86.7	89.6	97.7	98.2
12	86.5	89.5	97.4	98.2

Table 4
Influence of different distance measures for triplet sampling (Trip. Samp.) and similarity score (Sim. Scr.) measurement using our proposed **STCA** approach.

Trip. Samp.	Sim. Scr.	MARS		iLIDS-VIS
		mAP	rank-1	rank-1
Euclidean	Euclidean	80.6	85.5	85.3
Euclidean	Cosine	83.5	87.4	86.7
Cosine	Euclidean	84.8	89.2	87.3
Cosine	Cosine	**87.0**	**90.3**	**88.3**

the-art techniques irrespective of their backbone architectures in all datasets.

4.5. Ablation study

This section aims to show the effectiveness of each component of our proposed DL framework. First, we show the effect of varying resolutions with our proposed approach. Then, We study the impact of different sequence sizes and different distance measures with our proposed approach.

4.6. Influence of varying resolutions

To implement this experiment, we chose ResNet50-IBN [52] with the resolution of 256 × 128 and 384 × 128 on the MARS dataset. Here, we first reproduce the results of the *Baseline* [15] approach with both resolutions. Then, we did the experiments with cosine distance (CD) for triplet sampling. Finally, we conducted the experiments with all the components of the proposed framework with both resolutions.

Table 2 reported results of those experiments. By comparing *Baseline (2D-CNN or 3D-CNN)* to *Baseline (2D-CNN or 3D-CNN) + CD*, we can find significant improvement in accuracy with both resolutions. Including CD in the training process enables the model to converge faster (i.e. approximately 300 epochs faster). Results also suggest that simple *concatenation* of the feature representations from 2D-CNN and 3D-CNN does not help to improve the recognition accuracy. Including both 2D-CNN and 3D-CNN as proposed in our STCA framework improves the recognition accuracy significantly by the measures of 9.6% in mAP and 5.8% in rank-1 over the *Baseline (2D-CNN)* approach with 384 × 128 resolution. It also converges relatively faster than the other two alternatives. (See Table 3.)

In Fig. 5, we also demonstrate some qualitative examples from the MARS dataset which indicates that our proposed STCA approach effectively finds the true match in rank-1 (see the first row of each part of Fig. 5 (a)–(e)) when there are cases of misalignment, occlusions and body parts missing, while the 2D-CNN (Baseline) [15] approach finds it in later ranks. We also observed that there are few cases, as in Fig. 5 (e), where the true matches are only available in rank-1 using our STCA approach, while there are no true matches in the subsequent ranks. The qualitative analysis also suggests that there are a few cases where the proposed STCA approach is not able to find the true match in rank-01 (see the first row of Fig. 5 (f)) when all the images in a tracklet are well-aligned, although they are eventually recognized within the first few ranks (most cases in rank-2).

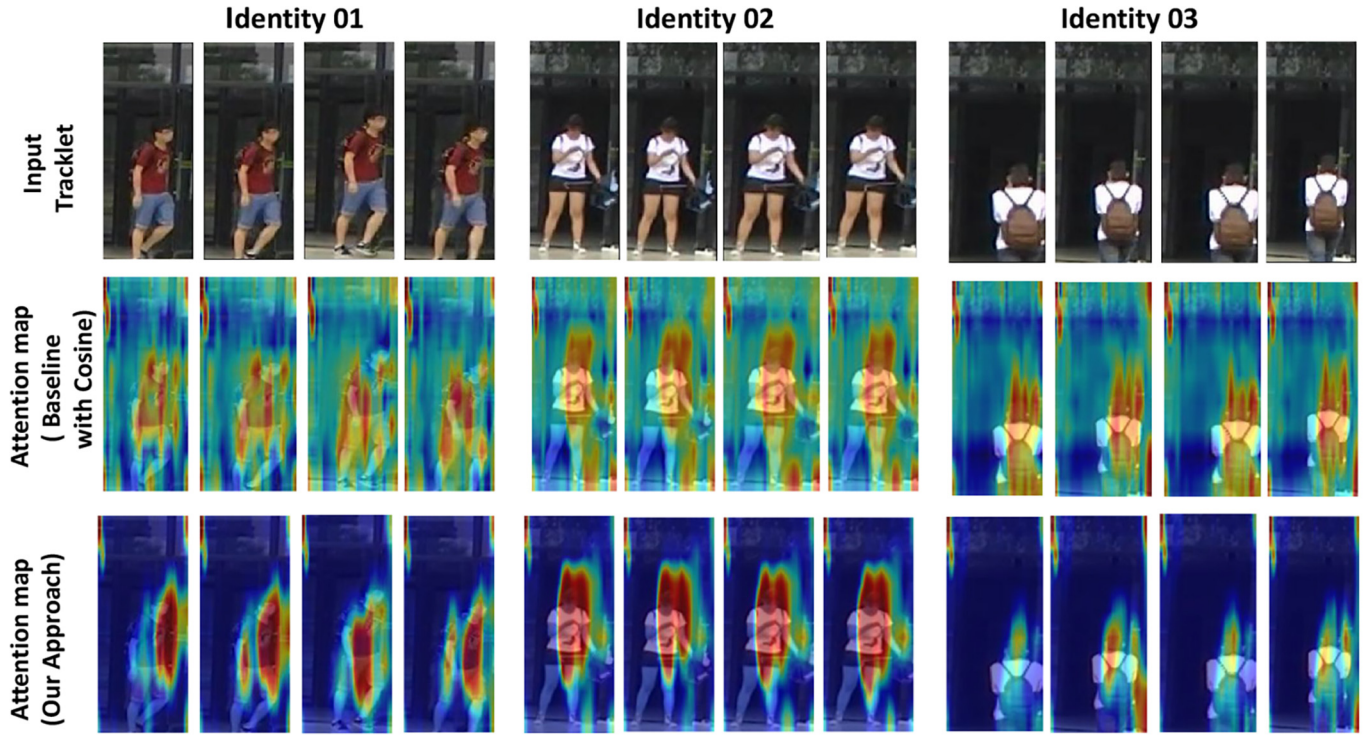

Fig. 4. Examples of tracklets of 3 individuals from the MARS dataset. The first row shows the original input tracklets; the second row shows the output attention map from the *Baseline (2D-CNN) with Cosine* approach, and the third row visualizes the attention map of our proposed STCA approach.

4.7. Influence of varying sequence size

In video ReID, the length of the sequence has an effect on the representative power of final aggregated feature embedding. Therefore, we did this experiment by considering the frame lengths 2,4,6,8,10,12, following the state-of-the-art [22,25] experimental setup. We found the optimal results by considering 8 frames in a sequence.

4.8. Influence of different distance measure

To implement this experiment, the proposed STCA approach is investigated with the resolution 384 × 128 on the MARS and iLIDS-VID dataset. We chose widely use Cosine and Euclidean distance measures, both for triplet sampling and measuring similarity scores. Fig. 6 presents the loss curves that have been while optimizing the STCA network using Cosine and Euclidean distance-based triplet sampling, respectively. As suggested in the reported curves, the lower-scale (i.e small angle) in cosine based hard triplet sampling as depicted in Fig. 6(a) and (b) prompted the network optimization to be converged faster than that of Euclidean distance based triplet sampling in Fig. 5 (c) and (d). To quantify this argument, we also reported the results in Table 4 for all combinations of Cosine and Euclidean measures. Reported results demonstrate that using Cosine measures for triplet sampling achieves a significant margin of improvement compared to utilizing Euclidean distance. We obtain the optimal performance for our approach when we used Cosine distance both for triplet sampling and measuring similarity scores.

4.9. Visualization

The goal of this section is to visualize the attention map with our proposed and the *2D-CNN + CD* approach. In this case, we select a few samples from particular tracklets with misalignment and visualize the attention map from the final 2D-CNN backbone layer. Fig. 4 evident that the attention map from our proposed STCA approach applies more weight (see. Third row) to the object of interest rather than scatter weight of the attention map with *2D-CNN + CD* approach.

4.10. Complexity analysis

Since the proposed STCA framework is the combination of 3D-CNN and 2D-CNN, the model complexity is higher in terms of the number of parameters and Multiply-Accumulate (MAC) operation. The proposed DL framework with *ResNet50-IBN* and *ResNet3D-50* has 41.87*M* parameters and 12.97*G* MACs operation. Mention that, the cross attention network itself has only .002514*M* parameters.

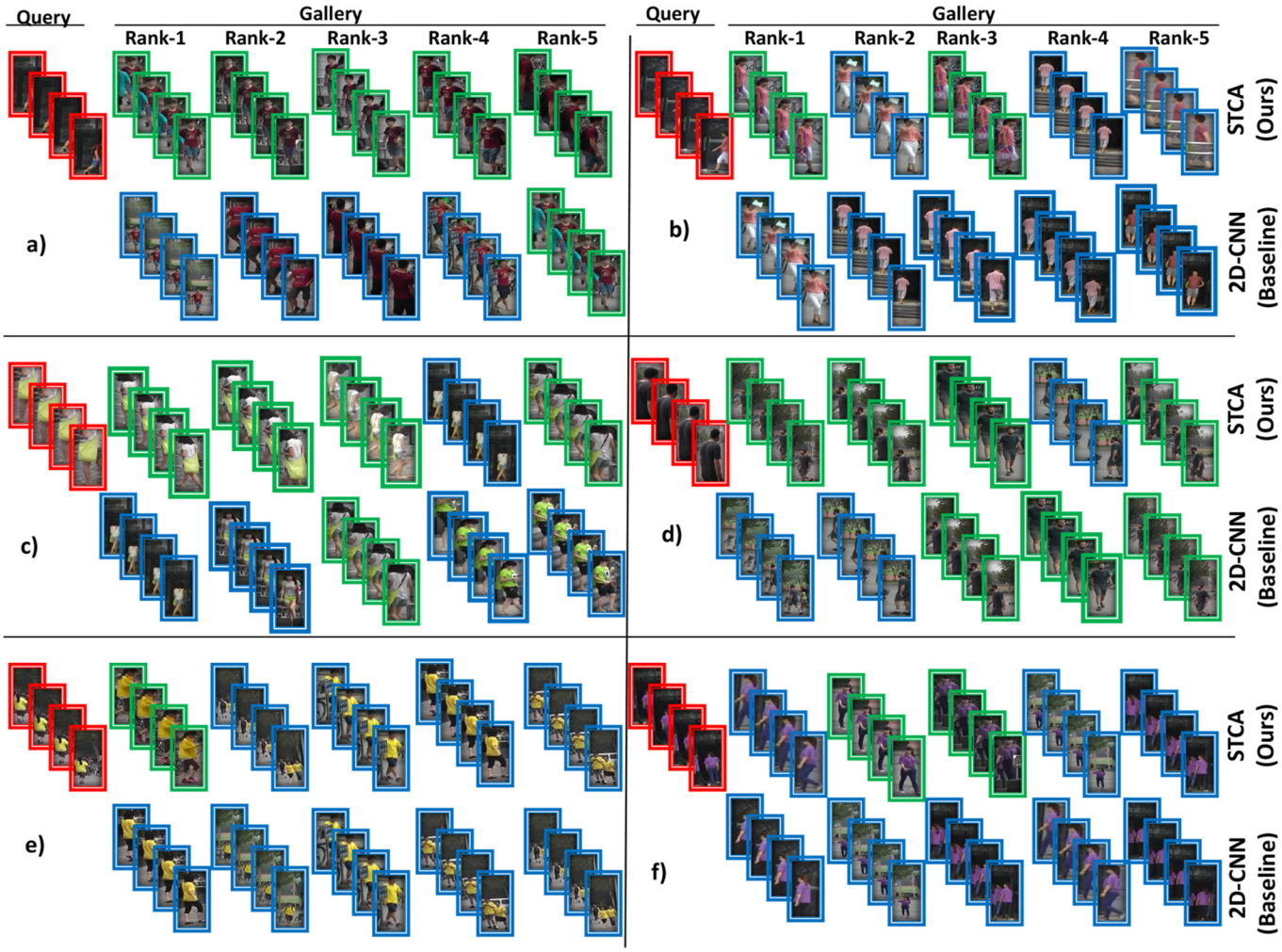

Fig. 5. Visual Comparison of query-set tracklets to top-5 matching tracklets for six random person in the MARS dataset. For each query, the first and second rows correspond to the ranking results produced by **STCA(Ours)** and **2D-CNN (Baseline)** approaches, respectively. Tracklets surrounded by green boxes denote the matches between query and gallery.

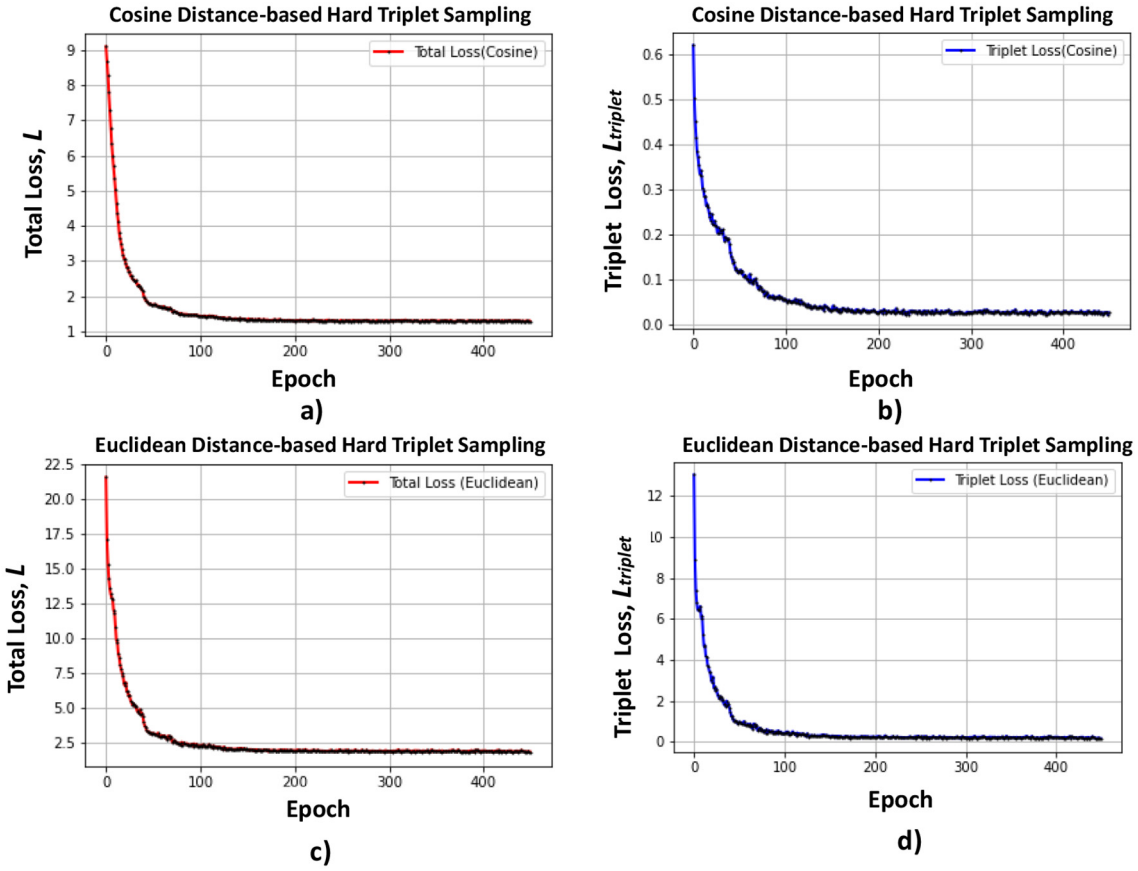

Fig. 6. STCA-based loss graphs for Euclidean and Cosine distance: (a) and (b) represent the total and triplet loss curves based on Cosine distance based hard triplet sampling, respectively; (c) and (d) represent the total and triplet loss curves based on Euclidean distance based hard triplet sampling, respectively.

5. Conclusions and future work

A novel DL framework (STCA) is proposed for video-based person ReID. A key component of this framework is the cross attention network including both 2D-CNN and 3D-CNN that dynamically selects the more relevant convolutional filters of the 2D-CNN backbone based on crossly attended spatio-temporal information, for enhanced feature representation and finer-grained inference. The proposed framework requires minimal modification to backbone 2D-CNNs commonly used for pairwise matching in ReID, and cross fusion can effectively combine spatio-temporal information to train the backbone network. Additionally, it exploits the benefits of using Cosine distance measure for batch-hard triplet mining. Experimental results on several benchmark datasets demonstrate that our proposed STCA framework can outperform related state-of-the-art methods, including approaches specialized to address misalignment issues.

To deal with the ReID challenges for gaining higher recognition accuracy, we proposed a DL framework STCA that has a greater computational complexity. Thus, future experiments could be exploring state-of-the-art pruning techniques guided by our proposed approach that could dynamically enable the pruning techniques to select and remove unnecessary filters, while trying to maintain a comparable accuracy.

Besides the pruning techniques, skipping the channels or frames through a learned decision policy could be beneficial to reduce the computational complexity for efficient video ReID. Low-level analysis of the STCA-based rank list shows that there is a number of missing true matches in rank-01. Thus, there is still room for improvement by utilizing a GAN-based sampling strategy for introducing synthetic im- ages in a batch while mining hard triplets for robust optimization and better recognition accuracy.

Credit authorship contribution statement

Amran Bhuiyan: Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Jimmy Huang:** Supervision, Investigation, Validation, Writing – review & editing.

Declaration of complete interest

The authors declare that they have no competing interests.

Acknowledgments

This research is supported by the research grant from Natural Sciences and Engineering Research Council (NSERC) of Canada and York Research Chairs (YRC) program. All the work was done when the first author was a postdoctoral fellow at the Information Retrieval and Knowledge Management Research Lab, York University, Canada. The authors gratefully appreciate the anonymous reviewers and associate editor for their valuable comments and constructive suggestions that greatly helped to improve the quality of the paper. We also acknowledge Compute Canada for providing us the computing resources to conduct experiments.

References

[1] Ejaz Ahmed, Michael Jones, Tim K. Marks, An improved deep learning architecture for person re-identification, In CVPR, 2015.
[2] Rodolfo Quispe, Helio Pedrini, Improved person re-identification based on saliency and semantic parsing with deep neural network models, Image Vis. Comput. 92 (2019), 103809.

[3] Khawar Islam, Person search: new paradigm of person re-identification: a survey and outlook of recent works, Image Vis. Comput. 101 (2020), 103970.

[4] Jingyi Lv, Zhiyong Li, Ke Nai, Ying Chen, Jin Yuan, Person re-identification with expanded neighborhoods distance re-ranking, Image Vis. Comput. 95 (2020), 103875.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, In CVPR, 2016.

[6] Alexander Hermans, Lucas Beyer, Bastian Leibe, In Defense of the Triplet Loss for Person Re-Identification. arXiv preprint, arXiv:1703.07737 2017.

[7] Kyoung Mu Lee, Part-aligned bilinear representations for person re-identification, in: Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei (Eds.),Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 402–419.

[8] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, Wei Jiang, Bag of tricks and a strong baseline for deep person re-identification, CVPRWK, 2019.

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

[10] Zhaohui Liang, Andrew Powell, Ilker Ersoy, Mahdieh Poostchi, Kamol- rat Silamut, Kannappan Palaniappan, Peng Guo, Md Amir Hossain, Antani Sameer, Richard James Maude, et al. Cnn-based image anal- ysis for malaria diagnosis. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 493–496. IEEE, 2016.

[11] Zhaohui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qm- ming Vivian Hu. Deep learning for healthcare decision making with emrs. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 556–559. IEEE, 2014.

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

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

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

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

[16] Fu Yang, Xiaoyang Wang, Yunchao Wei, Thomas Huang, Sta: Spatial-temporal attention for large-scale video-based person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8287–8294.

[17] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, Xilin Chen, Vrstc: Occlusion-free video person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 7183–7192.

[18] Xingyu Liao, Lingxiao He, Zhouwang Yang, Chi Zhang, Video-based person re-identification via 3d convolutional networks and non-local attention, Asian Conference on Computer Vision, Springer 2018, pp. 620–634.

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

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

[21] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Zhibo 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.

[22] Gu Xinqian, Hong Chang, Bingpeng Ma, Hongkai Zhang, Xilin Chen, Appearance-Preserving 3d Convolution for Video-Based Person Re-Identification, ECCV 2020.

[23] Madhu Kiran, Amran Bhuiyan, Louis-Antoine Blais-Morin, Ismail Ben Ayed, Eric Granger, et al., Flow guided mutual attention for person re-identification, Image Vis. Comput. 113 (2021) 104246.

[24] Fu Hui, Ke Zhang, Haoyu Li, Jingyu Wang, Zhen Wang, Spatial temporal and channel aware network for video-based person re-identification, Image Vis. Comput. (2021) 104356.

[25] Arulkumar Subramaniam, Athira Nambiar, Anurag Mittal, Co-segmentation inspired attention networks for video-based person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2019, pp. 562–572.

[26] Learning multi-granular hypergraphs for video-based person re-identification, in: Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, Ling Shao (Eds.),Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 2899–2908.

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

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

[29] Ruijie Quan, Yu Wu Xuanyi Dong, Linchao Zhu, Yi Yang, Auto-reid: Searching for a part-aware convnet for person re-identification, ICCV, 2019.

[30] Binghui Chen, Weihong Deng, Hu. Jiani, Mixed high-order attention network for person re-identification, ICCV, 2019.

[31] Meng Zheng, Srikrishna Karanam, Wu Ziyan, Richard J. Radke, Re-identification with consistent attentive siamese networks, CVPR, 2019.

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

[33] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Zhibo Chen, Densely semantically aligned person re-identification, CVPR, 2019.

[34] Djebril Mekhazni, Amran Bhuiyan, George Ekladious, Eric Granger, Unsupervised domain adaptation in the dissimilarity space for person re-identification, ECCV, 2020.

[35] Niall McLaughlin, Jesus Martinez Del Rincon, Paul Miller, Recurrent convolutional network for video-based person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 1325–1334.

[36] Jun Liu, Amir Shahroudy, Dong Xu, Gang Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, ECCV, 2016.

[37] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, Jiashi Feng, Video-based person re-identification with accumulative motion context, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2788–2802.

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He, Non-local neural networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 7794–7803.

[39] Xu Yiru Zhao, Zhongming Jin Shen, Hongtao Lu, Xian-sheng Hua, Attribute-driven feature disentangling and temporal aggregation for video person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 4913–4922.

[40] Dual attention matching network for context-aware feature sequence based person re-identification, in: Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, Gang Wang (Eds.),Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 5363–5372.

[41] Shuangjie Xu, Cheng Yu, Gu Kang, Yang Yang, Shiyu Chang, Zhou Pan, Jointly attentive spatial-temporal pooling networks for video-based person re-identification, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 4733–4742.

[42] Jiawei Liu, Zheng-Jun Zha, Wu Wei, Kecheng Zheng, Qibin Sun, Spatial-temporal correlation and topology learning for person re-identification in videos, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 4370–4379.

[43] Xian-Sheng Hua, Dense interaction learning for video-based person re-identification, in: Tianyu He, Xu Xin Jin, Jianqiang Huang Shen, Zhibo Chen (Eds.), Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, pp. 1490–1501.

[44] Fan Yang, Xiangtong Wang, Xuan Zhu, Binbin Liang, Wei Li, Relation-based global-partial feature learning network for video-based person re-identification, Neurocomputing 488 (2022) 424–435.

[45] Chenrui Zhang, Ping Chen, Tao Lei, Wu Yangxu, Hongying Meng, What-where-when attention network for video-based person re-identification, Neurocomputing 468 (2022) 33–47.

[46] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Yimin Liu, Jianguo Jiang, Saliency and granularity: discovering temporal coherence for video-based person re-identification, IEEE Trans. Circuits Syst. Video Technol. (2022).

[47] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K. Roy-Chowdhury, Wu. Ziyan, Spatio-temporal representation factorization for video-based person re-identification, Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, pp. 152–162.

[48] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, Xiaoyun Yang, Watching you: global-guided reciprocal learning for video-based person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 13334–13343.

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

[50] 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, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) July 2017, pp. 6776–6785.

[51] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint*, arXiv:1409.1556 2014.

[52] Xingang Pan, Ping Luo, Jianping Shi, Xiaoou Tang, Two at once: Enhancing learning and generalization capacities via ibn-net, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 464–479.

[53] Hu Jie, Li Shen, Gang Sun, Squeeze-and-excitation networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 7132–7141.

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the inception architecture for computer vision, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 2818–2826.

[55] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, Learning spatiotemporal features with 3d convolutional networks, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 4489–4497.

[56] Joao Carreira, Andrew Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 6299–6308.

[57] Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? Proceedings of the IEEE conference on Computer Vision and Pattern Recognition 2018, pp. 6546–6555.

[58] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, Xilin Chen, Cross attention network for few-shot classification, NeurIPS, 2019.

[59] Yandong Wen, Kaipeng Zhang, Zhifeng Li, Yu Qiao, A discriminative feature learning approach for deep face recognition, ECCV, 2016.

[60] Wang Cheng, Qian Zhang, Chang Huang, Wenyu Liu, Xinggang Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, In ECCV, 2018.

[61] Mars: A video benchmark for large-scale person re-identification, in: Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Su Chi, Shengjin Wang, Qi Tian (Eds.), European Conference on Computer Vision, Springer 2016, pp. 868–884.

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

[63] Taiqing Wang, Shaogang Gong, Xiatian Zhu, Shengjin Wang, Person re-identification by video ranking, European Conference on Computer Vision, Springer 2014, pp. 688–703.

[64] Olga Russakovsky, Jia Deng, Su Hao, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[65] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., The kinetics human action video dataset. arXiv preprint, arXiv:1705.06950 2017.

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

[67] Guiqing Liu, Wu. Jinzhao, Video-based person re-identification by intra-frame and inter-frame graph neural network, Image Vis. Comput. 106 (2021), 104068.

[68] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, Shiguang Shan, Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, pp. 2014–2023.

[69] Chanho Eom, Geon Lee, Junghyup Lee, Bumsub Ham, Video-based person re-identification with spatial and temporal memory networks, Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, pp. 12036–12045.

[70] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, Shiliang Zhang, In Proceedings of the IEEE International Conference on Computer Vision, 2019 3958–3967.