Video Person Re-identification with Competitive Snippet-similarity Aggregation and Co-attentive Snippet Embedding

Dapeng Chen[1,3] Hongsheng Li[1†], Tong Xiao[1] Shuai Yi[2] Xiaogang Wang[1†]
[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]SenseTime Research, [3]School of Software Engineering, Xi'an Jiaotong University
{dpchen, hsli, xiaotong, xgwang}@ee.cuhk.edu.hk yishuai@sensetime.com

Abstract

In this paper, we address video-based person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. Our approach divides long person sequences into multiple short video snippets and aggregates the top-ranked snippet similarities for sequence-similarity estimation. With this strategy, the intra-person visual variation of each sample could be minimized for similarity estimation, while the diverse appearance and temporal information are maintained. The snippet similarities are estimated by a deep neural network with a novel temporal co-attention for snippet embedding. The attention weights are obtained based on a query feature, which is learned from the whole probe snippet by an LSTM network, making the resulting embeddings less affected by noisy frames. The gallery snippet shares the same query feature with the probe snippet. Thus the embedding of gallery snippet can present more relevant features to compare with the probe snippet, yielding more accurate snippet similarity. Extensive ablation studies verify the effectiveness of competitive snippet-similarity aggregation as well as the temporal co-attentive embedding. Our method significantly outperforms the current state-of-the-art approaches on multiple datasets.

1. Introduction

Person re-identification (Re-ID) is a useful technology for intelligent video surveillance systems. Previous approaches mostly focus on image-based setting, where the trajectories of one person captured by different cameras are associated by comparing his/her still images. With the emergence of video-based benchmarks [27, 8, 34] and the growth of computational resource, researchers have started to utilize video data for the task. Video data contain much richer information about pedestrian appearance and also convey motion clues that implicitly reflect persons' body layouts. Both are beneficial and should be exploited for more robust re-identification.

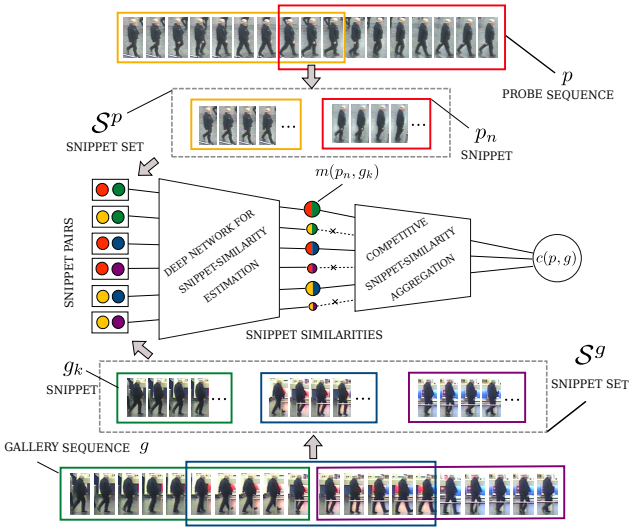


Figure 1: Illustration of the overall framework of our proposed approach. The long person sequences are divided into short snippets. Snippet similarities are robustly estimated by a deep neural network with a co-attentive mechanism. The sequence similarity competitively aggregates snippet-similarity scores.

For efficient Re-ID over a large-scale dataset, it requires learning a non-linear function to convert images/videos into a lower dimensional feature space. For video data, a straightforward solution is to embed the whole sequence into a single vector. However, as a person in one sequence may show great visual variation, a single vector cannot represent the diverse appearance of the person and will inevitably lose important details for identification. To solve this problem, we divide the sequence into multiple video snippets, embed each snippet separately, and perform snippet-similarity estimation and aggregation as in Fig. 1.

To obtain the similarity between two long sequences, the proposed competitive snippet-similarity aggregation

[†]H. Li and X. Wang are the co-corresponding authors.

scheme first breaks a long sequence into short snippets, then aggregates the top-ranked snippet-similarity scores for sequence-similarity estimation. With this scheme, the intra-sample visual variation could be minimized while the diverse appearance and temporal information of the sequence are maintained. The competitive aggregation scheme implicitly associates the snippets in one sequence with their most relevant snippets in the other one, yielding more accurate inter-sequence similarity.

The snippet similarity is estimated by a deep neural network with a co-attentive mechanism for snippet embedding. We employ multiple attention masks for different feature dimensions to softly select the features from different frames. The attention weights are obtained based on a query feature, which is learned from the whole probe snippet by an LSTM network. As it considers the overall snippet information, the resulting snippet embeddings are less affected by noisy frames caused by occlusion or low-quality detection. To better search the target person, the gallery snippets adopt the same query feature with the probe snippet, making the gallery snippet embeddings present more relevant features for snippet-similarity estimation.

Our main contributions are summarized as follows. (1) we propose a competitive similarity aggregation scheme with short snippet-based representations. It reduces the intra-person appearance variation for snippet-similarity estimation and aggregates diverse and reliable snippet-similarities to estimate the similarity between two sequences. (2) We propose a novel temporal co-attentive embedding for snippet-similarity estimation. It utilized a global query feature for the two compared images, which not only alleviates the influence of noisy frames but also guides more relevant feature embeddings for similarity estimation. (3) Our ablation studies validate the effectiveness of the proposed snippet competitive similarity aggregation and temporal co-attentive embedding. The final results surpass the previously best-published ones on three public benchmarks.

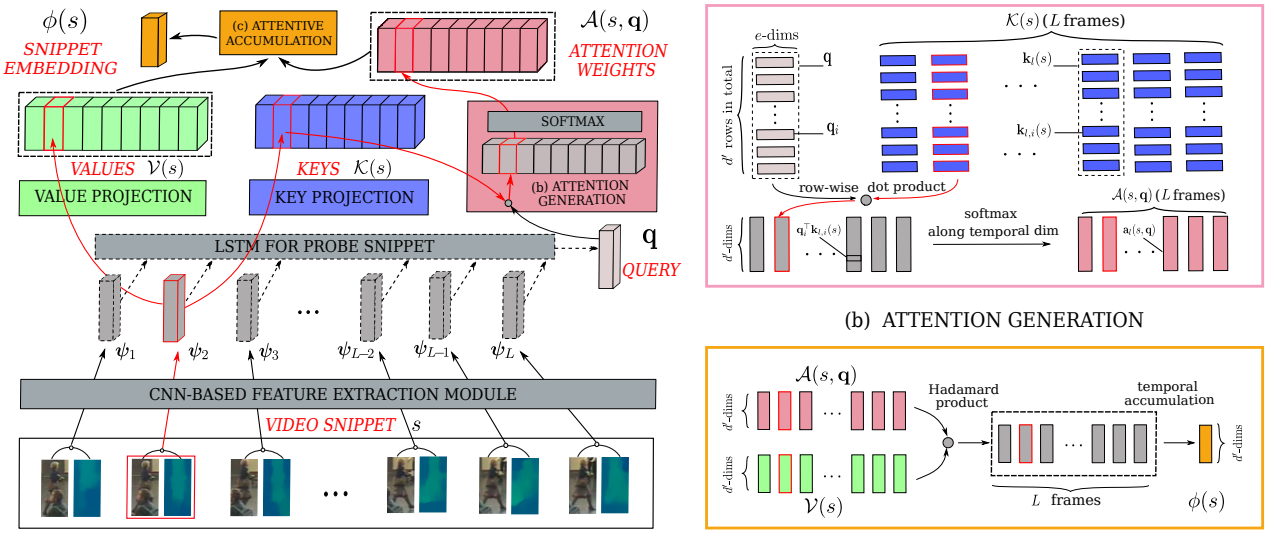
2. Related Work

Person re-identification has made significant strides. The researchers in recent years consider more realistic scenarios where the data size is much larger [35, 34], more complex [36, 21], and even in combination with other data modalities such as attributes or text descriptors [37, 12]. In particular, the video-based Re-ID setting is closer to practical scenarios as videos are the first-hand materials captured by surveillance cameras. They provide abundant appearance and motion information about persons and are promising for more accurate person re-identification [34, 27].

Early methods on video-based person Re-ID focus on handcrafting video representations. Wang *et al.* [27] combined HOG3D features and optical flow energy profile to obtain a spatiotemporal feature representation. Liu *et al.* [17] further incorporated spatial pictorial structures for spatially aligning person videos with different poses and viewpoints. Unsupervised embeddings like Fisher Vector [5] and Bag-of-Words [35] were adopted for encoding the low-level color and motion features. Metric learning algorithms were also developed for video-based Re-ID. Zhu *et al.* [40] and You *et al.* [31] imposed set-based constraints to better distinguish intra-person variations from the inter-person ones.

Since the breakthrough of deep Convolutional Neural Network (CNN) in image classification [10], researchers [13, 1, 22, 24, 16, 13, 11, 2, 33, 39, 32] have exploited the effectiveness of deep neural networks to learn more discriminative image representations from large-scale datasets. The powerful feature learning ability of CNN further facilitates the utilization of video data. McLaughlin *et al.* [19] built a CNN to extract features from each frame and then adopted a Recurrent Neural Network (RNN) to pass messages between different frames. Liu *et al.* [15] focused on learning motion context features from adjacent frames. Both methods adopted average/max pooling over the per-frame representations and outputted a single feature vector for the whole video. To better distill relevant information from the videos, very recent works introduced attention mechanisms to video-based person re-identification. Liu *et al.* [18] estimated a quality score for each frame to weaken the influence of the noisy samples. Zhou *et al.* [39] combined per-frame visual features and the forward propagated RNN hidden variables to generate the attention weights. Xu *et al.* [30] considered mutual influences between sequence pairs. The temporal attention weights for one sequence was guided by the features of the other sequence via a learned information sharing matrix.

Our proposed approach breaks a long person sequence into multiple short snippets, which is inspired by the success of video segments in action recognition [26, 25]. The sequence similarity is competitively aggregated over the top-ranked snippet similarities, achieving much better results than the methods encoding the whole video into a single feature vector [19, 15, 39, 30]. To measure the similarity between two snippets, our approach proposes to utilize a deep neural network with a temporal co-attentive mechanism for selecting representative frames and features from both snippets. Notably, our proposed attention is based upon "scaled dot-product attention" [23] in machine translation, which employs a query and a set of key-value pairs to generate the outputs. We have made two distinctive modifications over the previous one for video snippet embedding. (i) The query feature is learned from the whole probe snippet by an LSTM network thus is aware of the overall appearance and motion of the snippet. (ii) Multiple attention masks are generated, each of which softly accumulates the one-dimensional value features from different frames.



(a) TEMPORAL ATTENTIVE EMBEDDING

(b) ATTENTION GENERATION

(c) ATTENTIVE ACCUMULATION

Figure 2: Illustration of the proposed attention mechanism for snippet embedding. Both the probe snippet and gallery snippet adopt the same attention mechanism, but the query feature is from the compared probe snippet.

3. Our Approach

Video-based person Re-ID is essentially a ranking problem. Given a probe person sequence, the Re-ID task requires retrieving the same person's sequences from a gallery set and ranking them based on their similarities w.r.t. the probe sequence. To estimate the inter-sequence similarities, we break the long sequences into short snippets to first estimate the inter-snippet similarities, , then competitively aggregate the top-ranked snippet similarities (Section 3.1). The snippet similarities are estimated by a deep network with a co-attentive embedding strategy, which alleviates the influence of noisy frames and makes the embedding of a gallery snippet present more relevant information to compare with the probe snippet (Section 3.2). The overall framework is illustrated in Fig. 1.

3.1. Competitive Snippet-similarity Aggregation

Let p denote a probe sequence and g denote a gallery sequence, we aim to estimate the similarity between p and g. For this purpose, we generate a set of snippets from each sequence, where each snippet has a fixed length L and is sampled along the temporal dimension with a stride D. Suppose the probe sequence p has F^p frames, the snippets sampled from p form the set $\mathcal{S}^p = \{p_n\}_{n=1}^{N^p}$, where p_n is the nth snippet and N^p is the total number of snippets. When the frame number F^p is greater than L, we will have $N^p = \lfloor \frac{F^p - 1 - L}{D} \rfloor + 1$ snippets by discarding the several frames in the end. When F^p is fewer than L, we will generate only one snippet by utilizing all the frames and replicate the last frame until the snippet length achieves L.

The similarity between a probe sequence p and a gallery sequence g is calculated based on the similarities between the snippets from the two sequences. Denoting the similarity between snippets p_n and g_k by $m(p_n, g_k)$, where p_n and g_k are two arbitrary snippets in p and g. We can totally obtain $N^p \times N^g$ snippet-similarity scores between all possible snippet pairs in the snippet sets of the two sequences:

$$\mathcal{M}(p,g) = \{m(p_n, g_k) | p_n \in \mathcal{S}^p, g_k \in \mathcal{S}^g\}. \quad (1)$$

However, even for the sequences about one person, there might be many visual variations due to misalignments, occlusions, and different viewing angles. For two irrelevant snippets, such as one describing the front-view of a person and the other describing the back-view, the estimated similarity is less convincing. Instead of explicitly aligning the sequences in the temporal dimension [17], we propose a competitive strategy to solve the problem implicitly. Straightforwardly, we select the top-ranked snippet-similarity scores, *i.e.*, the highest $t\%$ values in $\mathcal{M}(p,g)$, to form the a more reliable similarity set $\widehat{\mathcal{M}}(p,g)$. Such selection is based upon the assumption that a higher similarity score indicates the compared two snippets being more relevant. The similarity between sequences p and g is then calculated as the average value of all scores in $\widehat{\mathcal{M}}(p,g)$, *i.e.*,

$$c(p,g) = \frac{1}{|\widehat{\mathcal{M}}(p,g)|} \sum_{m \in \widehat{\mathcal{M}}(p,g)} m. \quad (2)$$

3.2. Co-attentive Snippet Embedding

For estimating similarities for snippet pairs, we propose a novel deep neural network with a temporal co-attentive mechanism, which weights frame features according to

their importance and embeds a video snippet into a single vector. The similarity between the two snippets could then be estimated with the embedding vectors.

Given snippet s with L frames, we employ a visual CNN to extract the features from every frame. The feature vector of its lth frame is denoted by $\boldsymbol{\psi}_l(s)$, and the features of all the L frames are represented by the set $\boldsymbol{\Psi}(s) = \{\boldsymbol{\psi}_l(s)\}_{l=1}^{L}$.

3.2.1 Attention with Query and Key-value Features

For a video snippet s, $\boldsymbol{\Psi}(s)$ contains redundant information as there are only minor visual changes between neighboring frames. At the same time, the features of certain frames might be the outliers because of sudden occlusion or pedestrian mis-detection. To distill useful information from $\boldsymbol{\Psi}(s)$, we propose a novel attention mechanism to adaptively select the discriminative information from the sequentially varying features. Inspired by the "scaled dot-product attention" in [23], we produces a query feature for the whole snippet, at the same time, generate a key-value feature pairs for each frame in the snippet. The snippet embedding is a weighted summation of all frames' value features, where the weights are determined based on the compatibilities between the query feature and the key feature generated from that frame. Fig. 2a illustrates the flowchart of attentive embedding.

We build a value projection and a key projection for arbitrary frame l, which map $\boldsymbol{\psi}_l(s)$ to the key feature and value feature via linear projections, whose parameters are shared by all the frames. We compute these projections for $\boldsymbol{\psi}_l(s)$, feed the results to the corresponding batch normalization (BN) layer for normalization, and obtain the value feature $\mathbf{v}_l(s)$ and the key feature $\mathbf{k}_l(s)$. The value feature for the lth frame $\mathbf{v}_l(s)$ is a d'-dimensional vector, the key feature $\mathbf{k}_l(s)$ and the query feature $\mathbf{q}$ are $d'e$-dimensional vectors but are further reshaped to $d' \times e$ matrix for representation and computation convenience. Every e elements in $\mathbf{k}_l(s)$ and $\mathbf{q}$ are used to generate the weight for one element in $\mathbf{v}_l(s)$. The key features and value features of snippet s form the set $\mathcal{V}(s) = \{\mathbf{v}_l(s)\}_{l=1}^{L}$ and set $\mathcal{K}(s) = \{\mathbf{k}_l(s)\}_{l=1}^{L}$.

Different from other temporal attention mechanisms [39, 18] that only generate a single weight for the features in a frame, we aim to assign different weights to different elements in $\mathbf{v}_l(s)$. In this way, even if some elements in $\mathbf{v}_l(s)$ are contaminated, other elements can still be utilized for similarity estimation. With $\mathbf{k}_l(s)$ and $\mathbf{q}$, we generate d' temporal attention masks for one snippet. Specifically, let $\mathbf{k}_{l,i}(s) \in \mathbb{R}^e$ and $\mathbf{q}_i \in \mathbb{R}^e$ be the ith rows of $\mathbf{k}_l(s)$ and $\mathbf{q}$, the attention weighting value for $\mathbf{v}_{l,i}(s)$ is computed based on dot-product similarity between $\mathbf{q}_i$ and $\mathbf{k}_{l,i}(s)$ with a softmax non-linearization along the temporal dimension:

$$a_{l,i}(s, \mathbf{q}_i) = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_{l,i}(s))}{\sum_{l=1}^{L} \exp(\mathbf{q}_i^\top \mathbf{k}_{l,i}(s))}. \tag{3}$$

Based on Eq. (3), we obtain the attention weights $\mathbf{a}_l(s, \mathbf{q}) \in \mathbb{R}^{d'}$ for the lth frame. The attentions weights for all the frames in snippet s form the set $\mathcal{A}(s, \mathbf{q}) = \{\mathbf{a}_l(s, \mathbf{q})\}_{l=1}^{L}$. Fig. 2b illustrates the calculation.

3.2.2 Snippet Similarity with Co-attentive Embedding

To consistently embed a snippet pair, we utilize the "probe snippet" and the "gallery snippet" to indicate snippets from the probe sequences and the gallery sequences, respectively. The two kinds of snippets will play different roles in similarity estimation. Let p_n and g_k be an arbitrary probe snippet and a related gallery snippet, we expect the embedding of gallery snippet g_k can present more relevant information to the embedding of the probe snippet, in order to better measure their similarity.

We propose a temporal co-attention scheme where the value features in the probe snippet and gallery snippet are attended by the same query feature. To generate the query feature, the probe snippet p_n is processed by an LSTM, which consists of the following updating procedure:

$$\mathbf{h}_{l+1}(p_n) = LSTM\big(\boldsymbol{\psi}_l(p_n), \mathbf{h}_l(p_n)\big), \tag{4}$$

where the LSTM unit takes the single-frame visual features $\boldsymbol{\psi}_l(p_n)$ and the hidden states $\mathbf{h}_l(p_n)$ at current step as inputs and outputs hidden states of the next step. The hidden states at the last time step are summarization of overall appearance and motion information, which are used as the query feature for our attention model (Sec. 3.2.2):

$$\mathbf{q} = \mathbf{h}_L(p_n). \tag{5}$$

To embed the probe snippet p_n, we need to obtain the attention weights $\mathcal{A}(p_n, \mathbf{q})$ by the compatibilities between the query feature $\mathbf{q}$ and the key features in set $\mathcal{K}(p_n)$. The feature embedding of p_n is the summation of its per-frame value features $\mathbf{v}_l(p_n)$ weighted by the attention (Fig. 2c):

$$\phi(p_n) = \sum_{l=1}^{L} \mathbf{a}_l(p_n, \mathbf{q}) \circ \mathbf{v}_l(p_n), \tag{6}$$

where $\circ$ indicates the Hadamard product. To embed the gallery snippet g_k, the attention weights $\mathcal{A}(g_k, \mathbf{q})$ are calculated by the key features $\mathcal{K}(g_k)$ and the query feature $\mathbf{q}$. As $\mathbf{q}$ is generated from the probe snippet p_n, thus the embedding of g_k is dependent on p_n:

$$\phi(g_k|p_n) = \sum_{l=1}^{L} \mathbf{a}_l(g_k, \mathbf{q}) \circ \mathbf{v}_l(g_k). \tag{7}$$

The probe and the gallery snippets use the same query vector $\mathbf{q}$ to attend commonly interested information following Eqs. (6) and (7), making the similarity estimation between the two snippets more dependent on the probe person's characteristics.

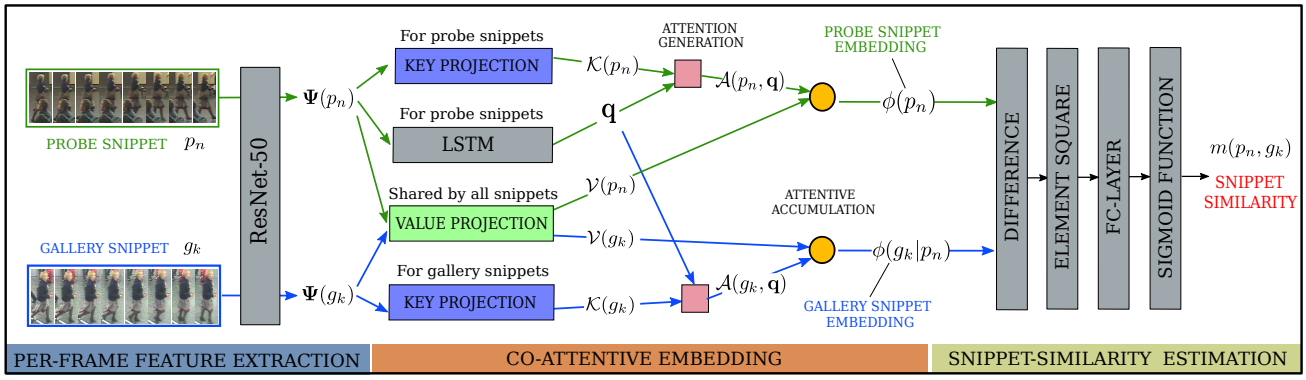


Figure 3: The network structure for pairwise snippet-similarity estimation. The probe snippet and gallery snippet have different pathways indicated by green and blue, respectively.

To estimate the snippet similarity $m(p_n, g_k)$ between p_n and g_k, we compute the difference vector of their feature embeddings $\phi(p_n)$ and $\phi(g_k|p_n)$, then perform element-wise square operation $(\cdot)^2$ over the difference vector. The resulting vector is transformed to a singular value by a fully-connected layer $f(\cdot)$, and the final similarity score is normalized to the range of $(0, 1)$ by a sigmoid function $\sigma(\cdot)$:

$$m(p_n, g_k) = \sigma(f((\phi(p_n) - \phi(g_k|p_n))^2)),$$

where $m(p_n, g_k)$ indicates the probability of p_n and g_k belonging to a same person.

3.2.3 The Network Structure

The structure of our snippet-similarity estimation network is illustrated in Fig. 3, which consists of a CNN module for learning per-frame visual feature, a co-attention module for snippet pair embedding and a similarity estimation module to output the snippet similarity. The network is trained in an end-to-end fashion.

Per-frame feature learning. We employ a CNN for obtaining visual features $\mathbf{\Psi}(s)$. The inputs to the network are the channel-wise concatenation of single-frame RGB and optical flow maps calculated by Epicflow [20], which are with the size of 128×64. The optical flow maps provide motion clues of the person's part or boundary location, thus is beneficial for identifying the person regions. We adopt the ImageNet [4] pretrained ResNet-50 [6], whose first conv-layer is modified to have 5 channels. A 2048-dimensional feature vector is extracted for each image as the output of the global average pooling layer after conv5_x [6].

Co-attentive embedding. Both value and key features are generated by a fully connected layer followed by a BN-layer [9] with the CNN features as their inputs. The query feature is based on the CNN features of the probe snippet by applying LSTM for summarization. The dimensionality of the value feature $\mathbf{v}_l(s)$, the key feature $\mathbf{k}_l(s)$ and the query feature $\mathbf{q}$ are 128, 128×4 and 128×4, respectively.

Snippet-similarity estimation. The module takes the probe snippet embedding $\phi(p_n)$ and gallery snippet embedding $\phi(g_k|p_n)$ as inputs. It then computes the element-wise square of the difference vector, projects it by a fully-connected layer, and finally outputs the probability of the two snippets being the same person.

3.2.4 Training Schemes

Loss functions. As $m(p_n, g_k) \in (0, 1)$, we can directly adopt the binary cross entropy loss to supervise the learning of snippet-similarity estimation. Define $m^*(p_n, g_k) = m(p_n, g_k)$ if sequences p and g belong to the same person, otherwise $m^*(p_n, g_k) = 1 - m(p_n, g_k)$, the loss function can be given as:

$$\mathcal{L}_{veri} = -\frac{1}{N_{veri}} \sum_{(p_n, g_k)} \log\left(m^*(p_n, g_k)\right), \qquad (8)$$

where N_{veri} is the number of sampled image pairs. Besides, we build a separated branch by taking the per-frame CNN features as input, which predicts person identities with supervision of the OIM (Online Instance Matching) loss [29],

$$\mathcal{L}_{id} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I} y^{i,n} \log\left(\frac{\exp(\mathbf{w}_i^\top x_n)}{\sum_{j=1}^{I} \exp(\mathbf{w}_j^\top x_n)}\right), \qquad (9)$$

where x_n indicates the CNN feature of nth image. The training set has N images belonging to I persons. If the nth image belongs to the ith person, $y^{i,n} = 1$, otherwise $y^{i,n} = 0$. $\mathbf{w}_i$ are the coefficients associated with the feature embedding of the ith person, which is online updated with the CNN feature of the ith person. They are obtained by using an online updated buffer and measuring similarities between the current person and all other persons in the feature buffer with inner product.

Network training. For network parameter training, we adopt stochastic gradient descent (SGD) with an initial learning rate of 10^{-3}, which is further decayed to 10^{-4} after the 20th epochs. To provide a reasonable number of positive and negative snippet pairs in each training mini-batch, we organize the data in the following way: the training videos are first divided into L-frame video snippets. Each batch contains the snippets from 32 random persons and each person has two randomly sampled snippets from different videos, where one snippet serves as the probe snippet and the other serves as the gallery snippet. There are 32 positive pairs and we randomly form 32×3 negative pairs with the 64 snippets, so the positive-to-negative rate is 1:3.

4. Experiments

We evaluate our proposed approach on three public datasets, iLIDS-VID [27], PRID-2011 [8] and MARS [34]. Ablation studies are conducted to investigate the effectiveness of the snippet-based representation, the competitive similarity aggregation and co-attentive snippet embedding from various aspects. The final performance of our approach outperforms those by state-of-the-art approaches.

4.1. Experimental Setup

Datasets. The iLIDS-VID dataset consists of 600 video sequences of 300 persons. Each sequence has a variable length ranging from 23 to 192 frames. This dataset is challenging because of high clothing similarities between different persons and the existence of occlusion. The PRID2011 dataset contains 749 persons, captured by two cameras, with sequence lengths ranging from 5 to 675 frames. The MARS dataset is a newly released dataset consisting of 1,261 pedestrians captured by at least 2 cameras. The bounding boxes are generated by classic detection and tracking algorithms, yielding 20,715 person sequences. Among them, 3,248 sequences are of quite poor quality due to the failure of detection or tracking, significantly increasing the difficulty of person Re-ID.

Experimental protocol and evaluation metrics. We follow the standard experimental protocols for testing on the datasets. For iLIDS-VID, the 600 video sequences of 300 persons are randomly split into 50% of persons for training and 50% of persons for testing. For PRID2011, we follow the experiment setup in previous methods [19, 27, 30, 39] and only use 400 video sequences of the first 200 persons, who appear in both cameras. The experiments on these two datasets are repeated 10 times with different test/train splits, and the results are averaged to ensure stable evaluation. For MARS, the predefined 8,298 sequences of 625 persons are used for training, while the 12,180 sequences of 636 persons are used for testing, including the 3,248 low quality sequences in the gallery set.

We use Cumulated Matching Characteristics (CMC) curve and mean average precision (mAP) to evaluate the performance for all the datasets. For ease of comparison, we only report the cumulated re-identification accuracy at selected ranks.

4.2. Analysis of Snippet Representation and Similarity Aggregation

We analyze the using of the snippet representation and also competitive similarity aggregation. Our standard version sets the snippet length $L = 8$, the stride $D = 4$, and the score selection rate $t\% = 20\%$.

Multiple snippets versus multi-shot images. One typical way to perform video-based Re-ID is to treat the sequence as a set of images and perform multi-shot image matching [18, 40, 31]. For a fair comparison, we design two multi-shot baselines reduced from our proposed framework. One handles the sequence in the form of individual frames (Table 1a), and the other further integrates the optical flow maps with RGB channels (Table 1b). Both of them learn to embed the frames rather than the snippets. Among the two baselines, incorporating optical flow can boost the top-1 accuracies by 2.9%, 3.6% and 4.3% on the three datasets, indicating that optical flow is valuable information for re-identification. Adopting the snippet representation (Table 1f) further improves the multi-shot baseline with optical flow by 18.7%, 6.7% and 6.4% in term of top-1 accuracy on iLIDS-VID, PRID2011 and MARS, showing that the snippet representation is the major factor that improves the multi-shot approaches.

Multiple snippets versus the complete sequence. We also design a variant of our framework that embeds the complete sequence (Table 1c) into a single feature vector. When testing with the complete sequence, we observe that our model trained with multiple snippets performs much better than the model trained with the entire sequences. This is because a person's video sequence might show much visual variations and cannot be effectively encoded by a single feature vector. The performance gap between this variant (Table 1 c) and our proposed approach (Table 1f) confirms our assumption.

Multiple snippets versus one snippet. To demonstrate the necessity of using multiple snippets for sequence-similarity estimation, we evaluate the single snippet-pair similarity. In testing, we utilize the same trained model as our standard one, but randomly sample one snippet to represent a sequence without the score aggregation procedure. The results in Table 1d significantly drop by 16.2%, 8.4% and 4.2% on top-1 accuracies compared with our standard version (Table 1f) for iLIDS-VID, PRID2011 and MARS, respectively.

The influence of the snippet length L . To investigate how

Tracklet representation	Optical flow	Network modules			iLIDS-VID			PRID2011			Mars		
		CNN	CE	SSE	top-1	top-5	mAP	top-1	top-5	mAP	top-1	top-5	mAP
a. multi-shot images	✗	✓	✗	✓	63.8	88.1	68.9	82.7	94.2	85.3	75.6	90.2	63.7
b. multi-shot images	✓	✓	✗	✓	66.7	89.8	71.3	86.3	96.2	88.5	79.9	91.6	66.6
c. complete sequence	✓	✓	✓	✓	74.4	92.5	78.4	86.0	97.8	88.7	82.4	92.9	67.5
d. one snippet (L=8)	✓	✓	✓	✓	69.2	90.9	73.8	84.6	96.6	87.2	82.1	93.4	71.1
e. multiple snippets (L=8)	✗	✓	✓	✓	79.8	91.8	82.6	88.6	99.1	90.9	81.2	92.1	69.4
f. multiple snippets (L=8)	✓	✓	✓	✓	**85.4**	**96.7**	**87.8**	**93.0**	**99.3**	**94.5**	**86.3**	**94.7**	**76.1**

Table 1: Comparison of different sequence representations with our proposed approach, where CNN, CE, SSE represent the CNN module for per-frame feature learning, temporal co-attentive embedding module and snippet-similarity estimation module, respectively. Top-1,-5 accuracies(%) and mAP (%) are reported.

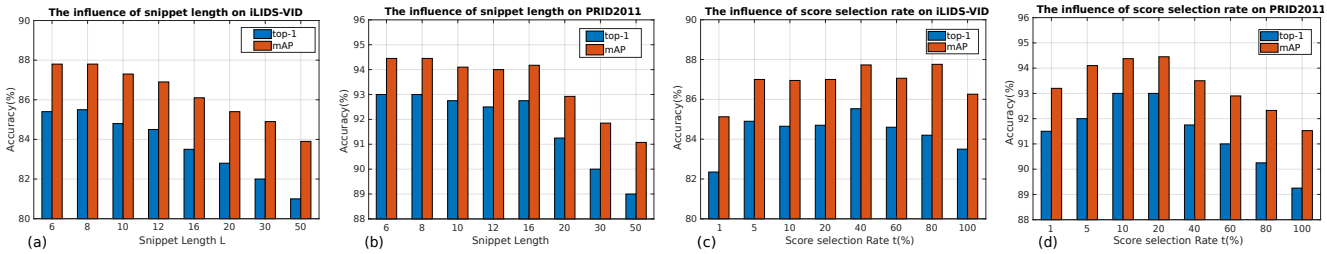


Figure 4: Parameter analysis for snippet length L and score selection rate $t\%$. (a)-(b) The top-1 accuracies and mAP on iLIDS-VID and PRID2011 with varying L. (c)-(d) The top-1 accuracies and mAP on iLIDS-VID and PRID2011 with varying $t\%$.

the snippet length influences the final performance, we conduct a series of experiments using snippets with different lengths for testing. In these experiments, when a sequence is shorter than the considered length, we use the whole sequence instead. Figs 4a and 4b show the top-1 accuracies and mAP on iLIDS-VID and PRID2011. It can be observed that the trained model is robust to the snippet length: all the top-1 accuracies are above 80% and the mAP are above 84%. We also find the results gradually decrease as the snippets become longer.

The influence of the score selection rate $t\%$. The competitive similarity aggregation strategy is expected to alleviate the effects of comparison between irrelevant snippets. However, too small $t\%$ might make the overall similarity estimation unstable. To investigate how $t\%$ influences the accuracy, we employ different score selection rates for testing. The results in Figs. 4c and 4d validate our assumption, where both too large or too small score selection rates lead to slightly worse performance, and the best results are achieved when $t\% = 20\%$.

4.3. Analysis of Co-attentive Snippet Embedding

We investigate the effectiveness of the proposed co-attentive embedding by comparing it with various temporal embedding strategies over the value features $\mathcal{V}(s)$.

Comparison of pooling strategies. Average/max pooling methods are the most straightforward ways to summarize the features from different frames. Based on snippet representations, these methods can yield reasonably well results as shown in Table 3a and 3b. However, they are less resistant to the contaminated frames in the snippet compared with the proposed co-attentive embedding (Table 3g).

Comparison with LSTM. LSTM is frequently adopted to encode a data sequence into a feature vector. We construct two variants based on LSTM: one encodes a snippet with its last hidden states (Table 3c), and the other utilizes the average vector of all the hidden states (Table 3d). In our evaluation, the effects of LSTM feature embeddings vary with different datasets. They are less effective than other temporal embeddings methods we studied here.

Comparison with different attention mechanisms. We also compare the proposed co-attentive embedding with two different attention mechanisms. (1) The first method is the self-attentive attention reduced from the proposed co-attention, where each snippet is encoded into a vector guided by the query feature learned from the snippet itself. (2) The second method was proposed in [39] for video person Re-ID. It utilizes the temporal varying LSTM hidden states to attend feature maps of the current frame in the sequence, which differs from the proposed attention mechanism of adopting the last hidden state as the only query feature to summarize both snippets. We adapt the attention mechanism in [39] to video snippets in our framework. Results in Table 3e and 3f suggest that our way of learning a global query feature vector is more effective than the temporal varying one for selecting representative features from different frames. The results in Table 3f and 3g confirm that the proposed co-attention mechanism is consistently better than the self-attentive version of our approach.

Methods	Deep model	iLIDS-VID				PRID2011			
		top-1	top-5	top-10	top-20	top-1	top-5	top-10	top-20
Wang *et al.* [28]	✗	39.5	61.1	71.7	81.0	40.0	71.7	84.5	92.2
Cho *et al.* [3]	✗	30.3	56.3	70.3	82.7	45.0	72.0	85.0	92.5
Mclaughlin *et al.* [19]	✓	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0
You *et al.* [31]	✗	56.3	87.6	95.6	98.3	56.7	80.0	87.6	93.6
Zheng *et al.* [34]	✓	53.0	81.4	- -	95.1	77.3	93.5	- -	99.3
Zhou *et al.* [39]	✓	55.2	86.5	- -	97.0	79.4	94.4	- -	99.3
Xu *et al.* [30]	✓	62.0	86.0	94.0	98.0	77.0	95.0	99.0	99.0
Liu *et.al.* [18]	✓	68.0	86.8	95.4	97.4	90.3	98.2	99.3	100.0
Proposed approach	✓	**85.4**	**96.7**	**98.8**	**99.5**	**93.0**	**99.3**	**100.0**	**100.0**

Table 2: Results by state-of-the-art methods on the iLIDS-VID and PRID2011 datasets. Top-1, -5, -10, -20 accuracies (%) are reported.

Methods	iLIDS-VID		PRID2011		MARS	
	top-1	mAP	top-1	mAP	top-1	mAP
a. ave-pool	79.8	82.6	87.8	90.2	83.7	75.9
b. max-pool	81.0	84.3	89.5	91.3	83.6	74.9
c. LSTM-last	78.0	82.0	83.3	86.2	84.4	73.9
d. LSTM-ave	72.0	76.6	87.5	89.8	82.6	71.9
e. att. in [39]	80.2	83.4	89.3	91.4	83.5	74.8
f. self-att.	84.7	87.4	89.6	91.6	84.8	75.5
g. co-att.	**85.4**	**87.8**	**93.0**	**94.5**	**86.3**	**76.1**

Table 3: Comparison of different snippet embedding methods. Top-1 accuracies (%) and mAP (%) are reported.

Methods	Deep model	MARS			
		top-1	top-5	top-20	mAP
Liao *et al.* [14]	✗	30.7	46.6	60.9	16.4
Li *et al.*[11]	✓	71.8	86.6	93.0	56.1
Zhou *et al.* [39]	✓	70.6	90.0	97.6	50.7
Liu *et al.* [18]	✓	73.7	84.9	91.6	51.7
Zhong *et al.*[38]	✓	73.9	- -	- -	68.5
Hermans *et al.* [7]	✓	79.8	91.4	- -	67.7
Proposed approach	✓	**86.3**	**94.7**	**98.2**	**76.1**

Table 4: Results by state-of-the-art methods on the MARS dataset. Top-1, -5, -20 accuracies (%) and mAP (%) are reported.

4.4. Comparison with State-of-the-art Approaches

We compare our approach with state-of-the-art approaches. The results are not refined by any post-processing techniques such as re-ranking [38] or multi-query [34].

Results on iLIDS-VID [27] and PRID2011 [8]. In Table 2, we compare our method with previous approaches that adopt various ways for video sequence embedding. Some of them adopt handcrafted features like color histogram [31, 3], HOG3D [31, 28]. Some others are based on deep learning frameworks, usually utilizing LSTM [19], average/max pooling [34] or temporal attention [39, 18] for feature summarization. Our approach significantly outperforms the others. The main reason for the improvements is that our method explicitly represents the sequences in the form of video snippets for both training and testing, while others either embed the whole sequence into a single vector [19, 31, 18, 39] or simply utilize individual images [3]. The proposed co-attention further makes the snippet embeddings more robust to outliers and more relevant to the probe snippet for similarity estimation.

Results on MARS [34]. MARS is currently the largest video person Re-ID dataset. Compared with iLIDS-VID, it is 4 times larger in the number of identities and 30 times larger in total sequences. For ease of training and testing, we randomly sample snippets from the sequences to keep the snippet number no more than 20 for each sequence. The results by our proposed method and state-of-the-art methods are shown in Table 4. The proposed approach significantly outperforms state-of-the-arts in terms of top-1, -5, -20 accuracies and mAP.

5. Conclusion

We proposed competitive similarity aggregation and co-attentive snippet embedding for the video-based person re-identification. Both strategies are based on video snippet representations, which reduce the intra-person variation in each sample and thus facilitate the similarity learning. The co-attentive snippet embedding alleviates the influences of noisy frames and enforces the compared snippet pairs weighting more on related information for snippet-similarity estimation. The competitive aggregation further employs reliable snippet similarities for final sequence-similarity estimation. We evaluate the proposed method on three datasets and perform a series of ablation studies to verify the effectiveness of each component of our approach. The final results are significantly better than those of the current state-of-the-art approaches.

Acknowledgement. This work is supported by SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong (Nos. CUHK14213616, CUHK14206114, CUHK14205615, CUHK14203015, CUHK14239816, CUHK419412, CUHK14207814, CUHK14208417, CUHK14202217), the Hong Kong Innovation and Technology Support Program (No.ITS/121/15FX), and China Postdoctoral Science Foundation (Nos. 2017M610641, 2014M552339).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2

[2] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017. 2

[3] Y.-J. Cho and K.-J. Yoon. Improving person re-identification via pose-aware multi-shot matching. In *CVPR*, 2016. 8

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 5

[5] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, and O. Camps. Person re-identification in appearance impaired scenarios. *arXiv preprint arXiv:1604.00367*, 2016. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[7] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 8

[8] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, 2011. 1, 6, 8

[9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *NIPS*. 2012. 2

[11] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 2, 8

[12] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *CVPR*, July 2017. 2

[13] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, June 2014. 2

[14] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 8

[15] H. Liu, Z. Jie, J. Karlekar, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *CoRR*, abs/1701.00193, 2017. 2

[16] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016. 2

[17] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *ICCV*, 2015. 2, 3

[18] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 2, 4, 6, 8

[19] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016. 2, 6, 8

[20] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *CVPR*, 2015. 5

[21] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. 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. 2

[22] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 4

[24] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 2

[25] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 2

[26] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[27] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014. 1, 2, 6, 8

[28] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, Dec 2016. 8

[29] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 5

[30] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, 2017. 2, 6, 8

[31] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, 2016. 2, 6, 8

[32] W. Zhang, X. Yu, and X. He. Learning bidirectional temporal cues for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2017. 2

[33] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 2

[34] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 1, 2, 6, 8

[35] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2

[36] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *CVPR*, 2017. 2

[37] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 2

[38] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 8

[39] 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 *CVPR*, 2017. 2, 4, 6, 7, 8

[40] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*, 2016. 2, 6