Local-Global Associative Frame Assemble in Video Re-ID

Qilei Li
q.li@qmul.ac.uk
Jiabo Huang
jiabo.huang@qmul.ac.uk
Shaogang Gong
s.gong@qmul.ac.uk

Computer Vision Group,
School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London E1 4NS, UK

Abstract

Noisy and unrepresentative frames in automatically generated object bounding boxes from video sequences cause significant challenges in learning discriminative representations in video re-identification (Re-ID). Most existing methods tackle this problem by assessing the importance of video frames according to either their local part alignments or global appearance correlations separately. However, given the diverse and unknown sources of noise which usually co-exist in captured video data, existing methods have not been effective satisfactorily. In this work, we explore jointly both local alignments and global correlations with further consideration of their mutual promotion/reinforcement so to better assemble complementary discriminative Re-ID information within all the relevant frames in video tracklets. Specifically, we concurrently optimise a local aligned quality (LAQ) module that distinguishes the quality of each frame based on local alignments, and a global correlated quality (GCQ) module that estimates global appearance correlations. With the help of a local-assembled global appearance prototype, we associate LAQ and GCQ to exploit their mutual complement. Extensive experiments demonstrate the superiority of the proposed model against state-of-the-art methods on five Re-ID benchmarks, including MARS, Duke-Video, Duke-SI, iLIDS-VID, and PRID2011.

1 Introduction

Person re-identification (Re-ID) aims to match pedestrian's identity across disjoint cameras views distributed at different locations [2, 34, 38, 39]. Early Re-ID studies concentrated on exploring appearance patterns unique per identity from still images [6, 21, 45], which has shown remarkable discrimination capacity. However, such methods assume well-curated data and the identity information are preserved in images. This assumption dramatically restricts their scalability and usability to many practical application scenarios when uncontrollable environments are the norm not the exception where video data are captured [20, 25]. Video person Re-ID beyond still images requires analysing and assembling information from a sequence of video frames in each tracklet so to build a more discriminative and robust representation of pedestrians in motion, minimising information corruption from poor frames and ID-switch [2, 4, 12, 24, 43, 46].

(a) Local part alignments (b) Global appearance correlations (c) Local-global Joint (d) Local-global Association

Figure 1: Illustration of four types of quality assessment strategies for frame assembling.

In the literature, one of the most commonly adopted techniques for assembling identity information from different video frames is *averaging* by pooling [30, 32]. By assuming all the frames are in equal importance, the pooling method neglects their diverse qualities caused by the constantly changing environments and/or unreliable pedestrian detections. Therefore, the aggregated tracket's representations are likely impacted by various types of noise as shown in Fig. 1. In order to *selectively* assemble video frames rather than averaging, attention mechanisms [13, 14, 33, 35, 37] have been studied to explore the correlations between the *global* visual features of frames (Fig. 1 (b)) so that the common appearance patterns shared among frames in the same tracklet are maintained while removing/ignoring unusual and low-quality frames [22, 23, 27, 36]. In contrast to the global appearance correlations, an alternative approach [11, 12, 48] compares video frames by *local* parts (Fig. 1 (a)) so to identify outliers that are significantly misaligned with other frames in a tracklet. Although sharing the same objective to adaptively assemble only the relevant video frames, these two approaches differ in exploiting information in different granularities. In isolation, both are sub-optimal in different real-world video scenes. The local-parts approach is fragile if the detected pedestrians are not well-aligned while the global-appearance approach is spatially insensitive, tending to miscorrelate patterns of interest in the background. Beyond attentive assembling, Recurrent Neural Network (RNN) [28, 43] has also been exploited for modelling temporal information to represent frame sequences in video tracklets. However, this approach is also vulnerable to noisy frames without careful frame selections [40].

In this work, we propose a tracklet frame assembling approach to video person Re-ID termed *Local-Global Associative Assembling* (LOGA). As shown in Fig. 1 (d), the LOGA method adaptively assembles video frames in the same tracklets by a Local Aligned Quality (LAQ) and a Global Correlated Quality (GCQ) modules to assess importance/relevance of the frames by both their alignments in local part and global appearance correlations as well as their mutual reinforcements. Whilst the focus of most existing spatial-temporal attentive methods is on collaborating the temporal information with *intra-frame* spatial attention, we aim to exploit the *inter-frame* complements more effectively, which is different and ready to benefit from the advancing per-frame learning. Specifically, the LAQ module divides all video frames in a tracklet into a same set of spatial parts and assesses each frame's quality by their part-wise alignment to the other frames so to measure both inter-frame visual similarity and spatial alignment. On the other hands, the GCQ module is applied on the holistic feature representation of each frame to consider inter-frame global appearance correlations, which is more robust to local part misalignment but spatially insensitive so less reliable from miscorrelation of information, *e.g.* irrelevant patterns in the background. Furthermore, to associate the local and global information and exploit their mutual benefits, we take the tracklet's representation assembled by the LAQ as its prototype and compare the global visual feature of frames with it in the GCQ module so that the two modules are encouraged to find a trade-off between the local and global information to cope with different types of noise more reliably.

Contributions of this work are three-fold: (1) To our best knowledge, we make the first attempt to explore the *association and mutual promotion* of frame's local part alignments

and global appearance correlations in assembling a sequence descriptor so to improve the model's robustness to noisy frames and inter-frame ID-switch in video Re-ID. (2) We propose a new video person Re-ID model termed *Local-Global Associative Assembling* (LOGA) that learns a discriminative and reliable representation for video tracklets by adaptively assembling frames of diverse qualities. (3) We introduce a local-assembled global appearance prototype to *associate* the local and global visual information by exploiting their mutual agreements to facilitate the learning of a discriminative tracklet representation.

Extensive experiments show the performance advantages and superior robustness of the proposed LOGA model over the state-of-the-art video Re-ID models on four video Re-ID benchmarks MARS [50], Duke-Video [29, 41], Duke-SI [20], and iLIDS-VID [34].

2 Related Works

Video person Re-ID aims to learn an expressive appearance feature and/or distance metric from a sequence of frames, *i.e*., a video tracklet. To take the advantages of the additional temporal information and complementary spatial information intrinsically available in video tracklets, existing approaches explore either local part alignments [1, 11, 12, 31, 48] or global appearance correlations [17, 18, 22, 23, 25, 27, 36, 47] to assemble the per-frame representations with high robustness to their diverse qualities.

Local part alignments. Considering the consistent body structure shared among humans and the arbitrary combinations of body part's appearance that unique to each identity, it is intuitive to differentiate images/frames of pedestrians regarding their visual similarity in different parts. In this spirit, local-parts assembling approaches [1, 11, 12, 31, 48] apply per-part comparisons of video frames in the same tracklets to identify outliers which are misaligned with others in most local parts, so as to restore the corrupted parts of frames with the complements of others [11, 12] or degrade their importance in frame assembling [1, 31, 48]. However, this hypothesis that a pedestrian detected in different video frames being mostly well-aligned is often untrue due to unreliable auto-generated person bounding boxes, *e.g*. the importance of a noise-free video frame might be underestimated due to the spatial shift of its detected bounding box from those in other frames. In this work, we further consider the holistic visual similarity of video frames when assessing their quality, which helps refrain from inaccurate assessments caused by part misalignments.

Global appearance correlations. In contrast to the local-parts approaches, methods based on global-appearance [17, 18, 22, 23, 25, 27, 36, 47] take the advantages of the strong representational power of convolutional neural network (CNN) [6, 16] to learn correlations between video frames holistically so that the irrelevant frames, which are likely in low-quality, are suppressed in frame assembling. However, the CNN features can be insensitive to spatial shift resulting in potential miscorrelations of visually similar but irrelevant parts, *e.g*. the ID-switch issue shown in Fig. 1 (b) is hard to be detected due to the subtle differences in the two pedestrians' outfits. This will result in misassemblling of frames to represent a tracklet. To address this problem, we propose to enhance the global-appearance methods by jointly explore frames' holistic visual correlations and their local part alignments by considering inter-frame spatial relations.

Spatial attention. Beyond the temporal assembling approaches discussed above, spatial attention [37] is also popular in both image and video person Re-ID [4, 21, 40, 42, 51]. By exploring the correlations of local parts within a still image or across different video frames, the spatial attention mechanism is able to adaptively focus on the more discriminative regions

regardless of their spatial location. However, this is prone to miscorrelation of information in video frames as in the global-correlated assembling approaches. Differently, our LAQ module investigates the alignments of the same part across different video frames, focusing on exploiting complementary inter-frame information in a tracklet.

There are a few recent attempts on exploring jointly the local and global information for frames assembling in video Re-ID [3, 44]. However, they learn from these two types of information with few interactions either by a dual-branch network [3] or feature concatenations [44], and overlook the local-global mutual impacts (Fig. 1 (c)). We validated the effectiveness of the proposed LOGA over those assembling strategies in both performance evaluation (Section 4.1) and ablation analysis (Section 4.2).

3 Video Person Re-ID

Given N video tracklets $\mathcal{T} = \{\boldsymbol{T}_i\}_{i=1}^N$ with each containing L frames $\boldsymbol{T}_i = \{\boldsymbol{I}_j^i\}_{j=1}^L$ depicting C pedestrians in motion, the objective of video person Re-ID is to derive a representation model θ from the tracklets data $\mathcal{V}$ which is capable of extracting discriminative feature representations $\boldsymbol{x}$: $f_\theta(\boldsymbol{T}) \rightarrow \boldsymbol{x}$ for Re-ID matching across disjoint camera views. Considering the diverse and unknown sources of noise commonly exist in surveillance videos, which leads to distractions in different frames, it is essential for the model to effectively recognise visual patterns that specific to each pedestrian to selectively assemble frames into a tracklet's representation. This is inherently challenging due to the uncertain nature of noise in tracklets of people in motion against backgrounds of visually similar distractors.

3.1 Local-Global Associative Assembling

In this work, we propose a *Local-Global Associative Assembling* (LOGA) model to address this problem by selecting information from video frames in the same tracklets according to both their local part alignments and global appearance correlations as well as the synergy and mutual promotion of these two types of information. For notation clarity, in the following, we focus on the formulation of assembling frames $\{\boldsymbol{I}_i\}_{i=1}^L$ in a single video tracklet $\boldsymbol{T}$ and ignore its tracklet index. As shown in Fig. 2, the video tracklet is first fed into a *Local Aligned Quality* (LAQ) module to assess the quality of frames regarding their part-wise alignment:

$$\{w_i^l\}_{i=1}^L = f_{\theta_l}(\{\boldsymbol{I}_i\}_{i=1}^L). \tag{1}$$

The θ_l in Eq. (1) is the learnable parameters of the LAQ and w_i^l denotes the importance of frames $\boldsymbol{I}_i$ determined by its alignments with other frames in local parts. Then, a *global correlated quality* (GCQ) module is devised which is applied to the D-dim holistic visual representation $\boldsymbol{E} = \{\boldsymbol{e}_i\}_{i=1}^L \in \mathbb{R}^{D\times L}$ of frames to determine their global appearance correlations. Instead of focusing on only the global visual features that are prone to spatial-insensitive miscorrelation, we explore the mutual synergy between local and global information by associating LAQ and GCQ through a prototypical descriptor $\boldsymbol{p}$. This assembles a frame's global features by their local-parts quality in GCQ for correlation exploration:

$$\boldsymbol{p} = \sum_{i=1}^{L} w_i^l \boldsymbol{e}_i, \tag{2}$$

$$\{w_i^g\}_{i=1}^L = f_{\theta_g}(\{\boldsymbol{e}_i\}_{i=1}^L | \boldsymbol{p}), \tag{3}$$

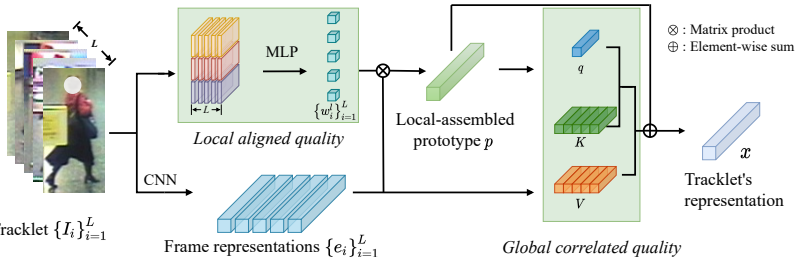


Figure 2: Overview of the proposed *Local-Global Associative Assembling* (LOGA) method.

where $\{w_i^g\}_{i=1}^L$ denotes frame's quality regarding their global-appearance feature $\boldsymbol{E}$ and θ_g is the learnable parameters of the GCQ module. In this way, the final representaion $\boldsymbol{x}$ of a tracklet $\boldsymbol{T}$ is obtained by associating LAQ and GCQ through $\boldsymbol{p}$:

$$\boldsymbol{x} = f(\boldsymbol{E}|\boldsymbol{w}^l; \boldsymbol{w}^g). \tag{4}$$

With the tracklet-level representations, the LOGA model can be trained with arbitarily conventional Re-ID objectives in an end-to-end manner. In inference, a generic distance metric (*e.g.* cosine distance) is used to measure pairwise visual similarity of tracklets for video Re-ID matching. The overall learning process of the LOGA model is depicted in Algorithm 1.

Algorithm 1 Local-Global Associative Assembling (LOGA).

Input: Video tracklets $\mathcal{T}$, Identity labels $\mathcal{Y}$.
Output: A deep CNN model for video person Re-ID.
for $i = 1$ **to** *max_iter* **do**
  Randomly sample a mini-batch of video tracklets from $\mathcal{T}$ and their identity labels from $\mathcal{Y}$.
  Compute the local-aligned per-frame importance scores (Eq. (1)).
  Feed the tracklets into backbone network to obtain their holistic visual features $\boldsymbol{E}$.
  Compute the local-assembled global appearance prototype (Eq. (2)).
  Compute the global-correlated per-frame importance scores (Eq. (3)).
  Compute the tracklet-level representations (Eq. (4)).
  Compute the objective losses and update the network by back-propagation.
end for

Local aligned quality. To explore the visual similarity of frames in terms of their local alignments, we separate them uniformly into M non-overlapping patches (parts) and apply patch-wise cross-frame convolution to recognise the aligned local patterns. This is accomplished by first flatten the 2D frames $\{\boldsymbol{I}_i\}_{i=1}^L$ then stacking them in the channel dimension as the raw representation of the tracklet $\boldsymbol{T}$ maintaining the inter-frames spatial correspondence. An 1D convolution is then applied on $\boldsymbol{T}$ to explore the per-part visual patterns,

$$\tilde{\boldsymbol{w}}^l = \boldsymbol{F} * \boldsymbol{T}, \quad \boldsymbol{F} \in \mathbb{R}^{S \times L \times L}, \tag{5}$$

where $*$ denotes the 1D convolution function and $\boldsymbol{F}$ is a trainable kernel. The size S of kernel $\boldsymbol{F}$ is determined by the granularity of the spatial separation, *i.e.*, $S = \frac{H \times W}{M}$ where H and W are the height and width of frames, respectively. The computed results $\tilde{\boldsymbol{w}}^l \in \mathbb{R}^{M \times L}$ encode

the part-wise importance of every frame, which is then aggregated by pooling followed by a multi-layer perceptron (MLP) to obtain the per-frame scores:

$$\boldsymbol{w}^l = f_{\theta_l}(\{\boldsymbol{I}_i\}_{i=1}^l) = \text{Softmax}(\text{MLP}(\text{Pooling}(\tilde{\boldsymbol{w}}^l))) \in (0,1)^{L\times 1}. \tag{6}$$

The Pooling$(\cdot)$ in Eq. (6) is a frame-wise mean pooling function and the MLP$(\cdot)$ stands for a single layer MLP activated by a ReLU function. The resulted scores are then normalised by softmax function as the indication $\boldsymbol{w}^l$ of per-frame importance to the tracklet $\boldsymbol{T}$. In this way, the LAQ learns to assess the frame's quality by its local part alignments to other frames, so to identify the misaligned outlier frames and suppress them from representing a tracklet.

Global correlated quality. The GCQ module is formulated to explore the inter-frame correlations according to their global appearances. However, the spatial invariant characteristic of the CNN features tends to miscorrelate patterns of interests with potential noise in the background, *i.e.* completely ignoring the spatial part's alignment. In this case, we propose to establish the GCQ on the results yielded by LAQ so to associate them by their synergy. Specifically, given the frame's importance $\boldsymbol{w}^l$ computed by Eq. (6) regarding their local part alignments, we first assemble their visual features accordingly in Eq. (2), which serves as the appearance prototype $\boldsymbol{p}$ of a tracklet. Then, the global-appearance quality of a frame is estimated according to the correlation between their global features and the prototype:

$$\begin{aligned} \boldsymbol{q} = f_{\theta_q}(\boldsymbol{p}) \in \mathbb{R}^{D\times 1}, \quad \boldsymbol{K} = f_{\theta_k}(\boldsymbol{E}) \in \mathbb{R}^{D\times L} \\ \boldsymbol{w}^g = f_{\theta_g}(\{\boldsymbol{e}_i\}_{i=1}^L | \boldsymbol{p}) = \text{Softmax}(\boldsymbol{K}^\top \boldsymbol{q}) \in (0,1)^{L\times 1}. \end{aligned} \tag{7}$$

The f_{θ_q} and f_{θ_k} functions in Eq. (7) are to linearly transform respectively the prototype and frame's features. Both are followed by batch normalisation. In this way, the video frames in $\boldsymbol{T}$ with higher appearance correlations to the pedestrian's prototype $\boldsymbol{p}$ will be highlighted with larger w_i^g and those mis-correlated ones will be suppressed.

Tracklet-level representation. Given the global-appearance quality of frames, their visual features can be selectively aggregated by:

$$\boldsymbol{V} = f_{\theta_v}(\boldsymbol{E}) \in \mathbb{R}^{D\times L}, \quad \hat{\boldsymbol{p}} = \boldsymbol{V}\boldsymbol{w}^g \in \mathbb{R}^{D\times 1}, \tag{8}$$

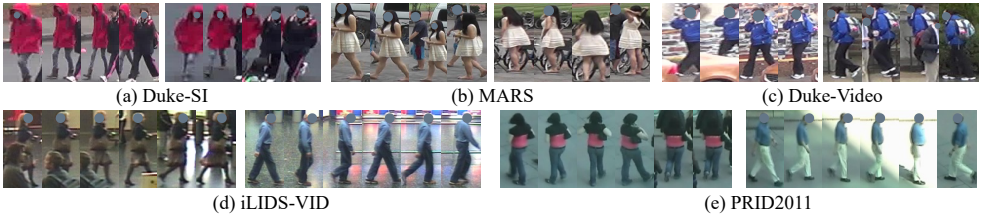
where f_{θ_v} is identical to f_{θ_q} and f_{θ_k} in Eq. (7) with independent parameters θ_v. Rather than taking $\hat{\boldsymbol{p}}$ as the final representation of the tracklet $\boldsymbol{T}$, in light of the residual learning [8], we distill the complementary information from global appearance correlations of frames to enhance the prototype computed by local-parts quality so to minimise representational error from identity-irrelevant part misalignments. To that end, we further learn the residual of $\boldsymbol{p}$ from $\hat{\boldsymbol{p}}$ and obtain the visual feature representation of $\boldsymbol{T}$ by:

$$\boldsymbol{x} = f(\boldsymbol{E}|\boldsymbol{w}^l; \boldsymbol{w}^g) = \boldsymbol{p} + \text{FC}(\hat{\boldsymbol{p}}) \in \mathbb{R}^{D\times 1}. \tag{9}$$

This design not only explores the global features of frames but also considers their local part alignments for optimising a discriminative tracklet representation.

3.2 Model Training

Given the formulations of LAQ and GCQ, the proposed LOGA model can benefit from conventional learning supervisions. Specifically, the LOGA model is jointly trained with a



(a) Duke-SI (b) MARS (c) Duke-Video

(d) iLIDS-VID (e) PRID2011

Figure 3: Example pairwise tracklets with the same ground-truth identity labels. Various noises are caused by illumination, viewpoints, resolution, occlusion, background clutter, etc.

softmax cross-entropy loss $\mathcal{L}_{\text{id}}$ and a triplet ranking loss $\mathcal{L}_{\text{trip}}$ [9]. The softmax cross-entropy loss $\mathcal{L}_{\text{id}}$ is employed to optimise identity classification:

$$\tilde{\boldsymbol{y}}_i = \text{Softmax}(\text{FC}(\boldsymbol{x}_i)), \quad \mathcal{L}_{\text{id}}(\boldsymbol{T}_i) = -\sum_{j=1}^{C} y_{i,j} \log \tilde{y}_{i,j}. \tag{10}$$

The $\boldsymbol{y}_i$ in Eq. (10) is an one-hot indicator of the ground-truth identity of tracklet $\boldsymbol{T}_i$ and the $\text{FC}(\cdot)$ serves as a linear classifier which maps the tracklet's representation $\boldsymbol{x}_i$ into an identity prediction distribution $\tilde{\boldsymbol{y}}_i$ while C is the total number of identities. Moreover, the triplet ranking loss $\mathcal{L}_{\text{trip}}$ explicitly draws the features of a positive tracklet pair sharing the same identity closer in the learned latent space while pushes the negative pairs apart:

$$\mathcal{L}_{\text{trip}}(\boldsymbol{T}_i) = \max(0, \Delta + \mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}_i^+) - \mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}_i^-)), \tag{11}$$

where $\boldsymbol{x}_i^+$ and $\boldsymbol{x}_i^-$ are the representations of two randomly sampled tracklets with the same and different ground-truth labels as $\boldsymbol{x}_i$ in respective, $\mathcal{D}(\cdot,\cdot)$ measures the distance of two features and Δ is a predefined margin. The overall optimisation objective of a batch of tracklets is then formulated by combining the two losses as:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} (\mathcal{L}_{\text{id}}(\boldsymbol{T}_i) + \mathcal{L}_{\text{trip}}(\boldsymbol{T}_i)), \tag{12}$$

where n is the size of a mini-batch. Since the objective function Eq. (12) is differentiable, the LOGA model can be trained end-to-end by the conventional stochastic gradient descent algorithm in the batch-wise manner.

4 Experiments

Datasets. The proposed Local-Global Associative Assembling (LOGA) is evaluated on four video-based Re-ID datasets: MARS [50], Duke-Video [29, 41], Duke-SI [20], iLIDS-VID [34], and PRID2011 [10]. Example tracklets are shown in Fig. 3. The MARS has 20,478 tracklets of 1,261 persons captured from a camera network with 6 near-synchronised cameras. Duke-Video is a newly released large-scale benchmark of 1,812 person identities with 4,832 tracklets. Duke-SI is a fully auto-generated version of Duke-Video without manual frames selection, thus, more practical and challenging. The iLIDS-VID dataset is relatively small scale including 600 video tracklets of 300 persons captured by two disjoint cameras in an airport arrival hall. The PRID2011 is another small scale dataset containing 1,134 tracklets from 934 identities captured by two cameras.

Evaluation Metrics. To evaluate the effectiveness of the proposed LOGA model, we adopted two commonly used performance metrics in person re-id including Cumulative Matching Characteristics (**CMC**) and Mean Average Precision (**mAP**) [49].

Implementation Details. For fair comparisons, we took a ResNet50 [8] as the backbone network for global visual feature extraction [7]. Given that the video tracklets are composed of arbitrary number of frames, we split each tracklet into several clips with a fixed length of 10. We randomly sampled 4 identity instances each with 8 clips to construct a mini-batch in model training. All the frames were resized to 256×128 and augmented by random horizontal flip. We used Adam [15] with weight decay of $5e-4$ for model optimisation. The margin Δ in Eq. (11) is set to 0.3, and the dimension D of representations is set to 2048 following [7, 26]. The kernel size S for the 1D convolution in Eq. (5) is set to 10. The model was trained on two P100 GPUs for 240 epochs, and the learning rate is initialised to $3e-4$ which linearly decayed with a factor of 0.1 per 60 training epochs. During the testing stage, the tracklet-level representation was obtained by averaging pooling the learned representations of their clips. Cosine distance was then used to measure the distances between a query and every probed tracklet in gallery for Re-ID.

4.1 Comparisons to the State-of-the-Art

Methods	Duke-Video				Duke-SI				MARS				iLIDS-VID			PRID2011		
	mAP	R1	R5	R20	mAP	R1	R5	R20	mAP	R1	R5	R20	R1	R5	R20	R1	R5	R20
TAUDL [23]	-	-	-	-	20.8	26.1	42.0	57.2	29.1	43.8	59.9	72.8	26.7	51.3	82.0	49.4	78.7	98.9
EUG [30]	78.3	83.6	94.6	97.6	-	-	-	-	67.4	80.8	92.1	96.1	-	-	-	-	-	
Snippet [3]	-	-	-	-	-	-	-	-	76.1	86.3	94.7	98.2	85.4	96.7	99.5	93.0	99.3	**100.0**
VRSTC [22]	93.5	95.0	99.1	99.4	-	-	-	-	82.3	88.5	96.5	-	83.4	95.5	99.5	-	-	-
GLTP [29]	93.7	96.3	99.3	99.7	-	-	-	-	78.5	87.0	95.8	98.2	86.0	98.0	-	95.5	**100.0**	-
UTAL [20]	-	-	-	-	36.6	43.8	62.8	76.5	35.2	49.9	66.4	77.8	35.1	59.0	83.8	54.7	83.1	96.2
STMP [22]	-	-	-	-	-	-	-	-	72.7	84.4	93.2	96.3	84.3	96.8	99.5	92.7	98.8	99.8
STA [4]	94.9	96.2	99.3	99.6	-	-	-	-	80.8	86.3	95.7	98.1	-	-	-	-	-	-
STAR [40]	93.4	94.0	99.0	99.7	-	-	-	-	76.0	85.4	95.4	97.3	85.9	97.1	99.7	93.4	98.3	**100.0**
FGRA [8]	-	-	-	-	-	-	-	-	81.2	87.3	96.0	98.1	88.0	96.7	99.3	95.5	**100.0**	**100.0**
MG-RAFA [36]	-	-	-	-	-	-	-	-	**85.9**	88.8	**97.0**	**98.5**	88.6	98.0	99.7	**95.9**	99.7	**100.0**
AP3D [6]	95.6	96.3	99.3	**99.9**	74.7	79.3	91.7	97.4	84.5	**90.4**	96.6	98.4	86.7	98.0	99.7	94.4	98.9	**100.0**
LOGA	**96.6**	**97.0**	**99.4**	**99.9**	**76.6**	**81.0**	**92.8**	**97.8**	84.1	89.5	96.3	97.9	**91.3**	**99.3**	**100.0**	**95.9**	98.9	**100.0**

Table 1: Comparisons to the state-of-the-art video person Re-ID methods. Results of the prior methods are from the original papers or reproduced by the official codes. The 1st/2nd best results are in **bold**/underlined. '†': unsupervised.

In Table 1, we compared the proposed LOGA model with a wide range of state-of-the-art video person Re-ID methods. The LOGA model yielded the best results across the board, which suggests the efficacy of associatively exploring local part alignments and global appearance correlation in assembling a discriminative representation of a tracklet. Whilst maintaining its competitiveness on the large-scale MARS and the well-curated Duke-Video datasets, the LOGA model achieved compelling improvements over the other methods on iLIDS-VID and its performance advantage is more significant on the automatically detected and segmented Duke-SI, in which case LOGA outperformed the others by 1.9%∼55%, 1.7%∼54.9% and 1.1%∼50% on mAP, rank-1 and rank-5, respectively.

4.2 Ablation Study

We conducted further studies to experimentally investigate the effectiveness of exploring the complementary local and global information by solely considering one while ablating another, and also demonstrated the superiority of our associative assembling over the dual-branch strategy [3] which used both local and global information separately. We also provided comprehensive visualisation for intuitively understandings.

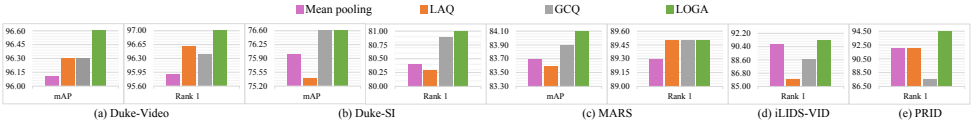


Figure 4: Ablation studies on model components.

Components analysis. We started with examining the role of local part alignments by introducing LAQ for frame assembling. Fig. 4 (pink *v.s.* orange) shows that both metrics on most datasets are decreased. This is caused by the unrealistic assumption that local regions of all the frames are well-aligned. Such an assumption is shown to be unreliable due to uncontrollable environment and fragile detection/segmentation. We further examined the importance of global appearance by solely employing GCQ for frame assembling. The unsatisfying performance as reported in Fig. 4 (pink *v.s.* gray) suggests assessing the quality of frames in accordance with solely the unobstructed global appearance is unreliable owing to the fine-grained details being ignored. In contrast, when both LAQ and GCQ are adopted, LOGA exhibits remarkable advantage over all other counterparts (green *v.s.* others). This demonstrates the indispensable of both LAQ and GCQ.

Effects of assembling strategy. We further studied the effects of different strategies to join the local and global information in frames assembling: (1) *separately* assembling by two individual branches learned in parallel according to the two kinds of information [3]. (2) *directly* connecting local and global information by rescaling the per-frame visual features $\boldsymbol{E}$ according to their normalised local alignment scores (Eq. (6)) then explore their global correlations by the conventional self-attention on the rescaled features. (3) *associatively* assembling by combining the local-assembled prototype and global-assembled residual (Eq. (9)) to exploit their synergy. The comparison given in Fig. 5 (green *v.s.* others) shows a noticeable advantage of LOGA over the dual-branch or direct-connecting counterpart, which demonstrates the effectiveness of the proposed associative assembling strategy.

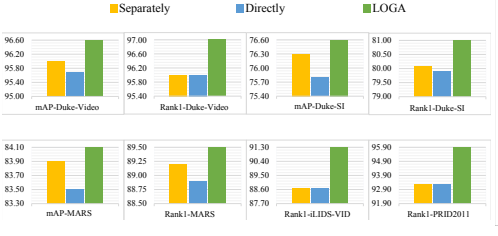


Figure 5: Impacts of assembling strategies.

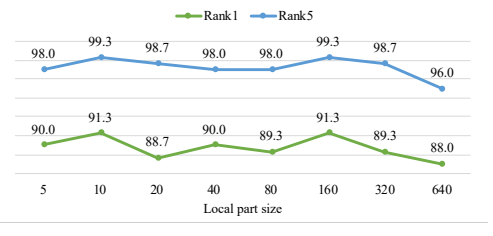


Figure 6: Impacts of local part size in LAQ.

Effects of local part size. We study the effects of local part size by varying the kernel size of the 1D convolution in Eq. (5) and experimented on iLIDS-VID. The experimental results shown in Fig. 6 indicates our model's robustness to this hyper-parameter within a wide range of values thanks to the subsequent GCQ module which help refine the local alignment scores according to global correlations. Given that improving S doesn't benefit the performance but increase the model's complexity, we set $S = 10$ in practice.

Qualitative studies. Fig. 7 shows several video clips stacked with their activation maps generated according to their local parts quality. Each frame's local-aligned score (upper, Eq. (6)) and global-correlated score (lower, Eq. (7)) are attached at their bottom-right corner.

As exhibited, LOGA is robust to various kinds of noise by providing a faithful importance score for assembling a discriminative representation. The activation maps accurately reveal the critical regions for Re-ID. The global-correlated scores are obtained with the complementary appearance information so can reliably adjust the biased local-aligned scores. For instance, as shown in Fig. 7, LAQ enables network to focus on the target instead of the switched ID or the irreverent multi-detected ID as shown in the activation maps. For the low quality frames caused by partial-detection, scale-variation and occlusion, etc. LAQ can faithfully assess the local quality. The suitable importance score revealed by the association of LAQ and GAQ efficiently guide LOGA to learn the representation from the most discriminative region in the most discriminative frames.

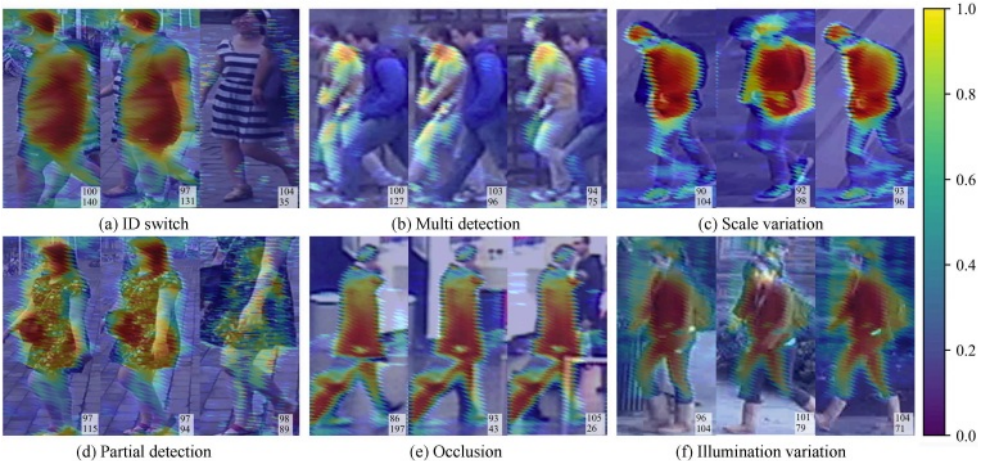


Figure 7: Visualisation of video clips suffering from various noise. Their corresponding importance in assembling are shown at the right-bottom corner of each frame with the local-alignment scores at top and the global-correlation scores beneath (amplified by 1,000 times).

5 Conclusions

In this work, we present a novel *Local-Global Associative Assembling* (LOGA) method for video person Re-ID through selectively assembling video frames of diverse qualities to derive a more reliable and discriminative representation of a video tracklet. This is accomplished by assessing the frame's quality according to both their *local part alignments* and *global appearance correlation* so to refrain from integrating undesired visual information into tracklet's representation causing identity mismatch. Different from existing approaches which explore either local or global information separately, our LOGA method constructs a local-assembled global appearance prototype of a tracklet so to alleviate biased quality assessment caused by either identity-irrelevant misalignment or spatial-insensitive appearance miscorrelation. Extensive experiments on five benchmark datasets show the performance advantages of LOGA over a wide range of the state-of-the-art video Re-ID methods. Detailed ablation studies are also conducted to provide in-depth discussions about the rationale and essence of different components in our model design.

References

[1] Liqiang Bao, Bingpeng Ma, Hong Chang, and Xilin Chen. Preserving structural relationships for person re-identification. In *ICMEW*, 2019.

[2] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, 2018.

[3] Zengqun Chen, Zhiheng Zhou, Junchu Huang, Pengyu Zhang, and Bo Li. Frame-guided region-aligned representation for video person re-identification. In *AAAI*, 2020.

[4] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, 2019.

[5] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press Cambridge, 2016.

[7] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, 2020.

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

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

[10] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[11] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, 2019.

[12] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, 2020.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[14] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, 2021.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436, 2015.

[17] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, 2019.

[18] Mengliu Li, Han Xu, Jinjun Wang, Wenpeng Li, and Yongli Sun. Temporal aggregation with clip-level attention for video-based person re-identification. In *WACV*, 2020.

[19] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018.

[20] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE TPAMI*, 42(7), 2019.

[21] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[22] Xingze Li, Wengang Zhou, Yun Zhou, and Houqiang Li. Relation-guided spatial attention and temporal refinement for video-based person re-identification. In *AAAI*, 2020.

[23] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. *CVPR*, 2021.

[24] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019.

[25] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017.

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

[27] Neeraj Matiyali and Gaurav Sharma. Video person re-identification using learned clip similarity aggregation. In *WACV*, 2020.

[28] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[30] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018.

[31] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018.

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

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2017.

[34] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[36] Zhikang Wang, Lihuo He, Xiaoguang Tu, Jian Zhao, Xinbo Gao, Shengmei Shen, and Jiashi Feng. Robust video-based person re-identification by hierarchical mining. *IEEE TCSVT*, 2021.

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[38] Guile Wu and Shaogang Gong. Decentralised learning from independent multi-domain labels for person re-identification. In *AAAI*, 2021.

[39] Guile Wu and Shaogang Gong. Generalising without forgetting for lifelong person re-identification. In *AAAI*, 2021.

[40] Guile Wu, Xiatian Zhu, and Shaogang Gong. Spatio-temporal associative representation for video person re-identification. In *BMVC*, 2019.

[41] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018.

[42] Wangmeng Xiang, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Part-aware attention network for person re-identification. In *ACCV*, 2020.

[43] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, 2017.

[44] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, 2020.

[45] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE TIP*, 2019.

[46] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *CVPR*, 2018.

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

[48] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, 2020.

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[50] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

[51] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *CVPR*, 2020.