Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-based Person Re-identification

Zhizheng Zhang[1*]   Cuiling Lan[2†]   Wenjun Zeng[2]   Zhibo Chen[1†]

[1]University of Science and Technology of China   [2]Microsoft Research Asia

zhizheng@mail.ustc.edu.cn   {culan,wezeng}@microsoft.com   chenzhibo@ustc.edu.cn

Abstract

*Video-based person re-identification (reID) aims at matching the same person across video clips. It is a challenging task due to the existence of redundancy among frames, newly revealed appearance, occlusion, and motion blurs. In this paper, we propose an attentive feature aggregation module, namely **M**ulti-**G**ranularity **R**eference-aided **A**ttentive **F**eature **A**ggregation (MG-RAFA), to delicately aggregate spatio-temporal features into a discriminative video-level feature representation. In order to determine the contribution/importance of a spatial-temporal feature node, we propose to learn the attention from a global view with convolutional operations. Specifically, we stack its relations, i.e., pairwise correlations with respect to a representative set of reference feature nodes (S-RFNs) that represents global video information, together with the feature itself to infer the attention. Moreover, to exploit the semantics of different levels, we propose to learn multi-granularity attentions based on the relations captured at different granularities. Extensive ablation studies demonstrate the effectiveness of our attentive feature aggregation module MG-RAFA. Our framework achieves the state-of-the-art performance on three benchmark datasets.*

1. Introduction

Person re-identification (reID) aims at matching persons in different positions, times, and camera views. Many researches focus on image-based setting by comparing the still images [42, 18, 12, 30, 16, 21]. With the prevalence of video capturing systems, person reID based on video offers larger capacity for achieving more robust performance. As illustrated in Figure 1, for a video clip, the visible contents of different frames differ but there are also overlaps/redundancy. In general, the multiple frames of a video clip/sequence can provide more comprehensive information
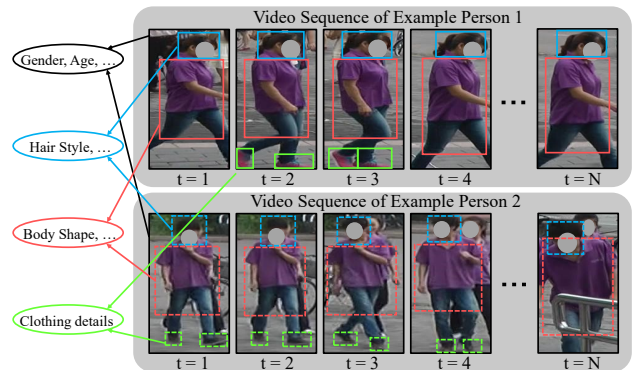

Figure 1. Illustration of two video sequences [1] of different identities. We observe that: (a) videos have redundancy with repetitive contents spanning over time; (b) there are some contents which occasionally appear but are discriminative factors (such as red shoes of person 1 in $t = 2, 3$). (c) discriminative factors/properties could be captured at different granularities/scales (*e.g.*, body shape can be captured from a large region (coarse granularity) while hair style is captured from small local region (fine granularity).

of a person for identification, but also raise more challenges, such as the handling of the presence of abundant redundancy, occlusion, motion blurs.

A typical video-based person reID pipeline [24, 20, 36, 6, 15, 5, 43, 22] extracts and aggregates spatial and temporal features to obtain a single feature vector as the video representation. To make the video-level feature representation precise, comprehensive and discriminative, we should latch onto the informative features from a global view, and meanwhile, remove the interference.

Attention, which aims at strengthening the important features while suppressing the irrelevant ones, matches the aforementioned goal well. Some works have studied spatio-temporal attention [15, 5] or the attentive recurrent neural networks [47, 21] to aggregate spatial and temporal features. They learn the attention weights for spatial and temporal dimensions separately or sequentially [47, 15]. However, due to the lack of a global view, they suffer from

---

[1]All faces in the images are masked for anonymization.

the difficulty in precisely determining whether a feature of some position is important and what the degree of redundancy is within the entire video clip. A diversity regularization is adopted in [15, 5] to remedy this issue, but only alleviates it to some extent. A powerful model is expected which jointly determines the importance levels of each spatio-temporal feature from a global view. Besides, as shown in Figure 1, discriminative factors/semantics could be captured at different granularities (regions of different sizes) by human. However, there is a lack of effective mechanisms to explore such hierarchical characteristics.

In this paper, we propose a Multi-Granularity Reference-aided Attentive Feature Aggregation (MG-RAFA) scheme for video-based person reID. For effective feature aggregation of the spatial and temporal positions, we determine the importance of each feature position/node from a global view, and consider the hierarchy of semantics during this process. For each feature position, we use its relations/affinities with respect to all reference feature nodes, which represent the global structural information (clustering-like patterns), together with the feature itself (appearance information) to model and infer the attention weights for aggregation. This is in part inspired by Relation-aware Global Attention (RGA) [41] designed for effective image feature extraction. However, a 3D video is rather different from a 2D image, where a video clip generally presents abundant redundancy along the time dimension, and the spatio-temporal structure patterns are complicated due to the diversity of human poses.

Considering the characteristics of video, we propose to construct a small but representative set of reference feature nodes (S-RFNs) for globally modelling the pairwise relations, instead of using all the original feature nodes. S-RFNs provides a simplified but representative reference for modeling global relations, which not only eases the difficulty of attention learning but also reduces the computational complexity. Moreover, we also take into account that the semantics are diverse in their granularities as illustrated in Figure 1. We propose to hierarchically model relations for attentive feature aggregation, which allows attention learning to be more precise and adaptive with low computational complexity.

In summary, we have three main contributions:

- For video-based person reID, we propose a simple yet effective Multi-Granularity Reference-Aided Attention Feature Aggregation (MG-RAFA) module for the joint spatial and temporal attentive feature aggregation.
- To better capture the discriminative semantics at different granularities, we exploit the relations at multiple granularities to infer attention for feature aggregation.
- We propose to build a small but representative reference set for more effective relation modeling by compressing the redundancy information of video data.

We conduct extensive experiments to evaluate our proposed feature aggregation for video-based person reID and demonstrate the effectiveness of each technical component. The final system significantly outperforms the state-of-the-art approaches on three benchmark datasets. Besides, the proposed multi-granularity module MG-RAFA further reduces the computational complexity as compared to its single-granularity version SG-RAFA via our innovative design. Our final scheme only slightly increases the computational complexity over the baseline ($< 1\%$).

2. Related Work

In many practical scenarios, video is ready for access and contains more comprehensive information than a single image. Video-based person reID offers larger optimization space for achieving high reID performance and attracts more and more interests in recent years.

Video-based Person ReID. Some works simply formulate the video-based person reID problem as an image-based reID problem, which extracts the feature representation for each frame and aggregates the representations of all the frames using temporal average pooling [27, 6]. McLaughlin *et al*. apply Recurrent Neural Network on the frame-wise features extracted from CNN to allow information to flow among different frames, and then temporally pool the output features to obtain the final feature representation [24]. Similarly, Yan *et al*. leverage LSTM network to aggregate the frame-wise features to obtain a sequence-level feature representation [37]. Liu *et al*. propose a two-stream network in which motion context together with appearance features are accumulated by recurrent neural network [19]. Inspired by the exploration of 3D Convolutional Neural Network for learning the spatial-temporal representation in other video-related tasks such as action recognition [11, 3], 3D convolution networks are used to extract sequence-level feature [17, 14]. These works treat the features with the same importance even though the features for different spatial and temporal positions have different contribution/importance levels for video-based person reID.

Attention for Image-based Person ReID. For image-based person reID, many attention mechanisms have been designed to emphasize important features and suppress irrelevant ones for obtaining discriminative features. Some works use the human part/pose/mask information to infer the attention regions for extracting part/foreground features [26, 12, 35, 26]. Some works learn the attention in terms of spatial positions or channels in end-to-end frameworks [18, 42, 16, 30, 38]. In [16], spatial attention and channel attention are adopted to modulate the features. In general, convolutional layers with limited receptive fields are used to learn the spatial attention. Zhang *et al*. propose a relation-aware global attention to globally learn attention by exploiting the pairwise relations [41] and achieve sig-

nificant improvement for image-based person reID. Despite the wide exploration in image-based reID, attention designs are under-explored for video-based reID, with much fewer efforts on the globally derived attention.

In this paper, motivated in part by [41] which is designed for effective feature learning of an image by exploring relations, we design a multi-granularity reference-aided attentive feature aggregation scheme for video-based person reID. Particularly, to compute the relations, we build a small set of reference nodes instead of using all the nodes for robustness and computational efficiency. Moreover, multi-granularity attention is designed to capture and explore the semantics of different levels.

Attention for Video-based Person ReID. For video-based person reID, some attention mechanisms have been designed. One category of works considers the mutual influences between the pair of sequences to be matched [36, 25, 4]. In [36], the temporal attention weights for one sequence was guided by the information from distance matching with the other sequence. However, given a sequence in the gallery set, it needs to prepare different features corresponding to different query images, which is complicated and less friendly in practical application.

Another category of works independently determines the features of the sequence itself. To weaken the influence of noisy frames, Liu *et al*. propose a quality aware network (QAN), which estimates the quality score of each frame for aggregating the temporal features as the final feature [20]. Zhou *et al*. use learned temporal attention weights to update the current frame features as input of RNN [47]. Zhao *et al*. disentangle the features of each frame to semantic attribute related sub-features and re-weight them by the confidence of attribute recognition for temporal aggregation [43]. These works do not simultaneously generate spatial and temporal attention from a global view for feature aggregation. Recently, spatial-temporal map which is directly calculated from the feature maps is used to aggregate the frame-level feature maps, without using any additional parameters [5]. However, as attention is computed in a predefined manner, the optimization space is limited.

Considering the lack of effective joint spatio-temporal attention mechanisms for the feature aggregation in video-based person reID, we intend to address this by proposing a multi-granularity reference-aided global attention, which jointly determines the spatial and temporal attention for feature aggregation.

3. Multi-Granularity Reference-aided Attentive Feature Aggregation

We propose an effective attention module, namely Multi-Granularity Reference-aided Global Attention (MG-RAFA), for spatial-temporal feature aggregation to obtain a video-level feature vector. In Section 3.1, we introduce

the preliminary. Then, we describe our proposed reference-aided attentive feature aggregation under single-granularity setting in Section 3.2 and elaborate the multi-granularity (our final design of MG-RAFA) in Section 3.3. We finally present the loss functions in Section 3.4.

3.1. Overview

For video-based person reID, we aim at designing an attentive feature aggregation module that can comprehensively capture discriminative information and exclude interference from a video which in general contains redundancy, newly revealed contents, occlusion and blurring. To achieve this goal, a joint determination of attention for the spatio-temporal features from a global view is important for robust performance.

We propose to learn the attention for each spatio-temporal position/node by exploring the global relations with respect to a set of reference feature nodes. Particularly, for the global relations modeling of a target node, we build a small set of representative feature nodes as reference, instead of using all the feature nodes, to ease optimization difficulty and reduce the computational complexity. Moreover, the discriminative information may physically spread over different semantic levels as illustrated in Figure 1. We thus introduce hierarchical (multi-granularity) relation modeling to capture the semantics at different granularities.

Figure 2 gives an illustration of our overall framework. For a video tracklet, we sample T frames as $\mathcal{V} = \{I_1, I_2, \cdots, I_T\}$. Through a single frame feature extractor (*e.g.*, ResNet-50 backbone), we obtain a set of feature maps $F_{all} = \{F_t | t = 1, 2, \cdots, T\}$, where $F_t \in \mathbb{R}^{H \times W \times C}$ includes $H \times W$ feature nodes, (H, W, C represent the height, width and number of channels, respectively). Based on the proposed multi-granularity reference-aided attention, all feature nodes in this set are weighted summed into a feature vector v as the final video-level feature representation for matching by l_2 distance. For clarity, we first present our proposed reference-aided attentive feature aggregation under the single-granularity setting in Subsection 3.2 and introduce the multi-granularity version in Subsection 3.3.

3.2. Reference-aided Attentive Feature Aggregation

The extracted feature set $F_{all} = \{F_t | t = 1, 2, \cdots, T\}$ consists of $K = H \times W \times T$ feature nodes, each of which is a C-dimensional feature vector. To determine the importance level of a feature node, it would be helpful if all the other feature nodes are also "seen", since intuitively people can determine the relative importance of something by comparing it with all others. For a feature node, to determine its importance level, we prepare its relation/affinity with every node as the ingredient to infer the attention. For any node i, when stacking its relations with all nodes (*e.g.*, in raster scan order), the number of relation elements is $D = H \times W \times T$.
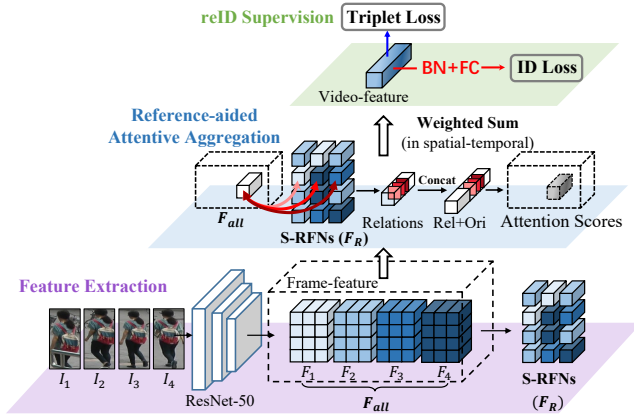
(a) Our pipeline with reference-aided attentive feature aggregation.
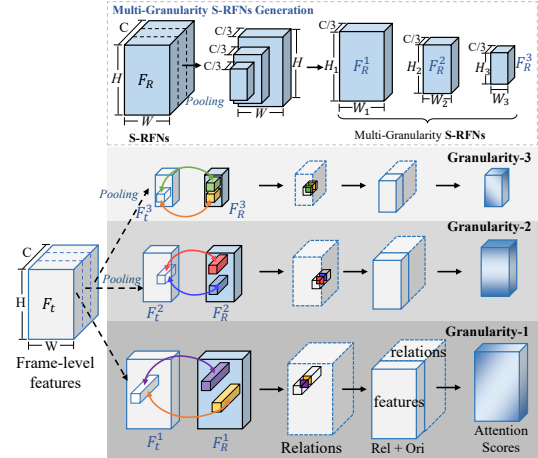
(b) Architecture of multi-granularity reference-aided attention.

Figure 2. Our proposed Multi-Granularity Reference-aided Attentive Feature Aggregation scheme for video-based person reID. (a) illustrates the reID pipeline with reference-aided attentive feature aggregation. Here, we use four frames ($T = 4$) as an example. For clarity, we only show the single-granularity setting in (a) and illustrate the procedure for deriving multiple-granularity reference-aided attention in (b). We use three granularity (N=3) here for illustration.

Taking into account the existence of appearance variations (*e.g.*, caused by pose, viewpoint variations) and large redundancy among frames, the distribution space for relation vectors is large and may cause difficulty in mining the patterns for accurately determining the attention. Thus, we propose to ease the difficulty by choosing a small set of representative feature nodes, instead of all the nodes, as the reference for modeling relations. As we know, for a video tracklet, there is usually large redundancy across temporal frames. For video action recognition, Bobick *et al.* [2] propose to use a static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location of a video sequence to compactly represent the information of a video. Motivated by this, we adopt average pooling along the temporal frames to fuse $F_{all} = \{F_t | t = 1, \cdots, T\}$ into a feature map $F_R \in \mathbb{R}^{H \times W \times C}$. Different from action recognition where the motion/temporal evolution is important, the temporal motion and evolution in general has no discriminative information for person ReID while the appearances are the key. Thus we simply average the temporal frames to obtain F_R as the reference, *i.e.*, the representative set of reference feature nodes (S-RFNs), to model global relations, which consists of $D = H \times W$ feature nodes.

For a feature node $\boldsymbol{x}^i \in \mathbb{R}^C$ in F_{all}, we calculate the relations/affinities between it and all feature nodes in the reference set F_R to model its corresponding relations. A pairwise relation is formulated as the correlation of the two nodes in embedding spaces:

$$r_{i,j} = r(\boldsymbol{x}^i, \boldsymbol{y}_r^j) = \mu(\boldsymbol{x}^i)^T \cdot \nu(\boldsymbol{y}_r^j), \quad (1)$$

where $\boldsymbol{y}_r^j \in \mathbb{R}^C$ denotes a feature node in the reference set F_R, i (and j) identifies the node index. We define

$\mu(\boldsymbol{x}^i) = ReLU(W_\mu \boldsymbol{x}^i)$ and $\nu(\boldsymbol{y}_r^j) = ReLU(W_\nu \boldsymbol{y}_r^j)$, where $W_\mu \in \mathbb{R}^{(C/s) \times C}$ and $W_\nu \in \mathbb{R}^{(C/s) \times C}$ are learned weight matrices, where s is a positive integral which controls the dimension reduction ratio. We implement it by adopting a 1×1 convolutional filter followed by Batch Normalization (BN) and ReLU activation, respectively. Note that we omit BN operations to simplify the notation. By stacking the pairwise relations of the feature node $\boldsymbol{x}^i$ with all the nodes (*e.g.*, scanned in raster scan order) in the reference set F_R, we have the relation vector as

$$\mathbf{r}_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,D}] \in \mathbb{R}^D, \quad (2)$$

which compactly reflects the global and clustering-like structural information. In addition, since the relations are stacked into a vector with a fixed scanning order with respect to the reference nodes, the spatial geometric information is also contained in the relation vector.

Intuitively, a person can have a sense of the importance levels of a node once he or she obtains the affinity/correlation of this node with many other ones. Similarly, the relation vector which describes the affinity/relation with all reference nodes provides valuable structural information. Particularly, the original feature $\boldsymbol{x}^i$ represents local appearance information while the relation feature $\boldsymbol{r}^i$ models global relations. They complement and reinforce each other but in different semantic spaces. We thereby combine them together in their respective embedding space and jointly learn, model, and inference the level of importance (attention scores) of the feature node $\boldsymbol{x}^i$ through a modeling function as

$$\boldsymbol{a}^i = \theta([\phi(\boldsymbol{x}^i), \psi(\boldsymbol{r}^i)]), \quad (3)$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are two embedding functions, $[\cdot, \cdot]$

represents the concatenation operation, and $\theta(\cdot)$ represents a modeling function to inference the attention vector $\boldsymbol{a}^i \in \mathbb{R}^C$ corresponding to $\boldsymbol{x}^i$. We define $\phi(\boldsymbol{x}^i) = ReLU(W_\phi \boldsymbol{x}^i)$, $\psi(\boldsymbol{r}^i) = ReLU(W_\psi \boldsymbol{r}^i)$, and $\theta([\phi(\boldsymbol{x}^i), \psi(\boldsymbol{r}^i)]) = ReLU(W_\theta([\phi(\boldsymbol{x}^i), \psi(\boldsymbol{r}^i)]))$, where $W_\phi \in \mathbb{R}^{(C/s) \times C}$, $W_\psi \in \mathbb{R}^{(D/s) \times D}$ and $W_\theta \in \mathbb{R}^{C \times ((C/s) + (D/s))}$ are learned weight matrices. We implement them by performing 1×1 convolutional filtering followed by BN and ReLU. For each feature node $\boldsymbol{x}^i$ in F_{all} (nodes corresponding to all the spatial and temporal positions), we obtain an attention score vector $\boldsymbol{a}^i$. For all nodes in F_{all}, we have $\mathcal{A} = [\boldsymbol{a}^1, \boldsymbol{a}^2, \cdots, \boldsymbol{a}^K]$.

We normalize the learned attention scores via the Softmax function across different spatial and temporal positions (node indexes) and obtain the final attention $\hat{\boldsymbol{a}}^i$, $i = 1, 2, \cdots, K$. Afterwards, we use the final attention as the weights to aggregate all the feature nodes (from all spatial and temporal positions) in F_{all}. Mathematically, we obtain the final sequence-level feature representation $\boldsymbol{v} \in \mathbb{R}^C$ by

$$\boldsymbol{v} = \sum_{k=1}^K \hat{\boldsymbol{a}}^i \odot \boldsymbol{x}^i, \quad \hat{\boldsymbol{a}}^i = Softmax(\boldsymbol{a}^i), \qquad (4)$$

where symbol $\odot$ represents element-wise multiplication.

3.3. Multi-Granularity Attention

Human can capture the different semantics (such as gender, body shape, clothing details) of a person at different granularity levels (*e.g.*, in terms of viewing distance or image resolution). Some types of semantics (*e.g.*, whether the person wears glasses or not) may be easier to capture at fine granularity while some others (*e.g.*, body shape) may be easier to capture at coarse granularity by excluding the distraction from fine details. Motivated by this, we propose the Multi-Granularity Reference-aided Attentive Feature Aggregation (MG-RAFA) which derives the attention and introduces a hierarchical design, aiming at capturing the discriminative spatial and temporal information at different semantic levels. Basically, we distinguish the different granularities by modeling relations and deriving attention on feature maps of different resolutions.

Following the earlier notations, for both reference nodes in F_R and the nodes to be aggregated in F_{all}, we split them along their channel dimensions into N splits/groups. Each group corresponds to a granularity. In this way, we reduce the computational complexity in comparison with the single granularity case. For the mth granularity, we perform spatial average pooling with a ratio factor m on both the mth split features of F_R and F_t, $t = 1, 2, \cdots, T$. We obtain the factorized reference feature $F_{R,m} \in \mathbb{R}^{H_m \times W_m \times \frac{C}{N}}$ of $D_m = H_m \times W_m$ nodes, where $H_m = \frac{H}{2^{m-1}}$ and $W_m = \frac{W}{2^{m-1}}$. Similarly, we obtain the factorized feature map on frame t as $F_{t,m} \in \mathbb{R}^{H_m \times W_m \times \frac{C}{N}}$ and the spatial and tempo-

ral feature node set as $F_{all,m} = \{F_{t,m} | t = 1, 2, \cdots, T\}$.

Then, we employ the reference-aided attentive feature aggregation as described in Section 3.2 for each group separately. Thereby, the relation modeling in Eq. (1) and the attention modeling function in Eq. (3) can be expanded into their multi-granularity versions as

$$r(\boldsymbol{x}_m^i, \boldsymbol{y}_m^j) = \mu_m(\boldsymbol{x}_m^i)^T \cdot \nu_m(\boldsymbol{y}_m^j), \qquad (5)$$

$$\boldsymbol{a}_m^i = \theta_m([\phi_m(\boldsymbol{x}_m^i), \psi_m(\boldsymbol{r}_m^i)]), \qquad (6)$$

where subscript m identifies the index of granularity, $\boldsymbol{x}_m^i$ denotes the ith node in $F_{all,m}$ and $\boldsymbol{y}_m^j$ denotes the jth nodes in the reference feature map $F_{R,m}$. Similar to the feature aggregation under single granularity in Section 3.2, we normalize the attention scores via Softmax function and weighted sum the feature nodes (across different spatial-temporal positions). Finally, we concatenate the fused feature of each split/group (denoted by $\boldsymbol{v}_m$) to obtain the final sequence-level feature representation $\boldsymbol{v} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N]$.

3.4. Loss Design

We add the retrieval-based loss, *i.e.*, the triplet loss with hard mining L_{Tr}, and the ID/classification loss (cross entropy loss with label smoothing [29]) denoted by L_{ID}, on the video feature vector $\boldsymbol{v}$. Each classifier consists of a Batch Normalization (BN) layer followed by a fully-connected (FC) layer. Specially, to encourage the network to aggregate discriminative features at each granularity, we add the two loss functions on the aggregated feature of each granularity $\boldsymbol{v_g}, g = 1, \cdots, N$. The final loss is:

$$L = L_{ID}(\boldsymbol{v}) + L_{Tr}(\boldsymbol{v}) + \frac{1}{N} \sum_{g=1}^N (L_{ID}(\boldsymbol{v}_g) + L_{Tr}(\boldsymbol{v}_g)). \quad (7)$$

4. Experiments

4.1. Datasets and Evaluation Metrics

Table 1. Three public datasets for video-based person reID.

Datasets	MARS [44]	iLIDS-VID [31]	PRID2011 [9]
Identities	1261	300	200
Tracklets	20751	600	400
Distractors	3248 tracklets	0	0
Cameras	6	2	2
Resolution	128×256	64×128	64×128
Box Type	detected	manual	manual
Evaluation	CMC & mAP	CMC	CMC

We evaluate our approach on three video-based person reID datasets, including MARS [44], iLIDS-VID [31], and

Table 2. The ablation study for our proposed multi-granularity reference-aided global attention (MG-RAFA) module. Here, "SG" denotes "Single-Granularity" and "MG" denotes "Multi-Granularity". N denotes the number of granularities. S denotes the number of splits(groups) along the channel dimension for masking attention on each split respectively. In a multi-granularity setting, the number of splits is equal to the number of granularities (*i.e.*, $S = N$) since each split correponds to a granularity level. We use "MG-AFA" to represent the attention module without relations, in which attention values are inferred from RGB information alone.

Models	#GFLOPs	Mars				iLIDS-VID			PRID2011		
		mAP	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20
Baseline	32.694	82.1	85.9	95.1	97.3	86.5	96.6	98.9	92.5	98.5	99.6
MG-AFA (*N=4*)	+0.095	82.5	86.6	96.1	97.8	86.7	96.6	98.7	92.6	98.1	99.6
SG-RAFA (*S=1*)	+2.301	85.1	87.8	96.1	**98.6**	87.1	97.1	99.0	93.6	98.2	99.9
SG-RAFA (*S=4*)	+0.615	84.9	88.4	96.6	98.5	86.7	96.6	98.7	94.2	98.6	99.6
MG-RAFA (*N=2*)	+0.742	85.5	88.4	**97.1**	98.5	87.1	97.3	99.3	94.2	98.2	99.9
MG-RAFA (*N=4*)	**+0.212**	**85.9**	**88.8**	97.0	98.5	**88.6**	**98.0**	**99.7**	**95.9**	**99.7**	**100**

PRID2011 [9]. Table 1 gives detailed information. Following the common practices, we adopt the Cumulative Matching Characteristic (CMC) at Rank-1 (R-1), to Rank-20 (R-20), and the mean average precision (mAP) as the evaluation metrics. For MARS, we use the train/test split protocol defined in [44]. For iLIDS-VID and PRID2011, similar to [19, 4, 15], we report the average CMC across 10 random half-half train/test splits for stable comparison.

4.2. Implementation Details

Networks. Similar to [1, 23, 5], we take ResNet-50 [7] as our backbone for per-frame feature extraction. Similar to [28, 39], we remove the last spatial down-sampling operation in the conv5_x block for both the baseline and our schemes. In our scheme, we apply our propose MG-RAFA after the last residual block (conv5_x) for the attentive feature aggregation to obtain the final feature vector v. We build *Baseline* by taking the feature vector v obtained through global spatial-temporal average pooling.

Experimental Settings. We uniformly split the entire video into $T=8$ chunks and randomly sample one frame per chunk. For the triplet loss with hard mining [8], in a mini-batch, we sample P identities and each identity includes Z different video tracklets. For MARS, we set $P=16$ and $Z=4$ so that the mini-batch size is 64. For iLIDS-VID and PRID2011, we set $P=8$ and $Z=4$. We use the commonly used data augmentation strategies of random cropping [33], horizontal flipping, and random erasing [46, 33, 30] (with a probability of 0.5 *at the sequence level* for both the baselines and our schemes. Sequence-level data augmentation is much superior to frame-level one. This is closer to the realistic data variation and will not break the inherent consistency among frames. We set the input resolution of images to be 256×128 with $T=8$ frames. Adam optimizer is used. Please find more details in the supplementary.

4.3. Ablation Study

4.3.1 Effectiveness Analysis

We validate the effectiveness of our proposed multi-granularity reference-aided attention (MG-RAFA) module

and show the comparisons in Table 2.

MG-RAFA vs. Baseline. Our final scheme *MG-RAFA* (*N = 4*) outperforms *Baseline* by **2.9%**, **2.1%** and **3.4%** on Mars, iLIDS-VID, and PRID2011, respectively. This demonstrates the effectiveness of our proposed attentive aggregation approach.

Effectiveness of using the Global Relations. In Table 2, *MG-AFA (N=4)* denotes the scheme when we use the visual feature alone (without using relations) to learn the attention. Our scheme *MG-RAFA (N=4)* which uses relations outperforms *MG-AFA (N=4)* by 2.2%, 1.9%, and 3.3% in Rank-1 on Mars, iLIDS-VID, and PRID2011, respectively, indicating the effectiveness of leveraging global relations for learning attention.

Single Granularity vs. Multiple Granularity. *SG-RAFA(S=1)* also takes advantage of relations but ignores the exploration of semantics at *different granularities*. In comparison with *SG-RAFA(S=1)*, our final scheme *MG-RAFA(N=4)* that explores relations at multiple granularities achieves 1.0%, 1.5%, 2.3% Rank-1 improvements on Mars, iLIDS-VID, PRID2011, respectively. *MG-RAFA* is effective in capturing correlations of different granularity levels.

Moreover, we study the effects of different numbers of granularities by comparing *MG-RAFA(N=4)* with *MG-RAFA(N=2)*. The results show that finer granularity delivers better performance. Note that the spatial resolution of the frame features F_t is 16×8. The spatial resolution ratio between two adjacent granularity levels is set to 4, which allows the maximum number of granularity levels N to be 4 (*i.e.*, 16×8, 8×4, 4×2, and 2×1) in this work. In the subsequent description, we use *MG-RAFA* to refer to *MG-RAFA(N=4)* unless otherwise specified.

To further demonstrate that the improvements come from the relation modeling at varying granularities instead of multiple attention masks, we compare *MG-RAFA(N=4)* with the single-granularity setting *SG-RAFA(S=4)*. In this setting, the features are divided into four splits(groups) along channels with each split having an attention mask rather than a shared one. Each attention mask is derived from the same fine gruanularity. The results show that *MG-*

Table 3. Comparison of different strategies on selection of the reference feature nodes (S-RFNs). Different spatial(S) and temporal (T) pooling strategies are compared. We denote the spatial and temporal node dimension as $(H \times W \times T)$. For example, we build the S-RFNs by adopting temporal average pooling, leading to $16 \times 8 \times 1$ nodes in S-RFNs.

S-RFNs	#Nodes	#GFLOPs	Mars				
			mAP	R-1	R-5	R-10	R-20
Baseline	0	32.694	82.1	85.9	95.1	96.5	97.3
S-P (8×1×8)	64	+2.034	83.9	86.6	96.1	97.4	98.0
S-P (8×2×8)	128	+2.345	83.9	87.2	95.6	97.2	97.9
S-P (4×4×8)	128	+2.345	84.1	87.1	95.7	97.3	97.9
S-P (8×4×8)	256	+2.967	84.2	87.0	95.8	97.4	98.0
T-P (16×8×2)	256	+2.916	84.7	87.4	96.1	97.4	98.3
T-P (16×8×4)	512	+4.159	84.7	87.3	96.1	97.4	98.1
ST (16×8×8)	1024	+6.697	84.3	87.3	95.8	97.2	98.1
Ours (16×8×1)	128	+2.301	**85.1**	**87.8**	**96.1**	**97.8**	**98.5**

RAFA(N=4) is superior to *SG-RAFA(S=4)*.

Complexity. Thanks to the channel splitting and spatial pooling, the computational complexity (FLOPS) of our multi-granularity module *MG-RAFA(N=4)* is only 9.2% of that of the single granularity module *SG-RAFA(S=1)*.

4.3.2 Selection of the Reference Feature Nodes

In our scheme, we take a set of feature nodes (S-RFNs) as reference to model pairwise relations. Instead of taking all feature nodes in the frame features as reference, considering the larger temporal redundancy, an average pooling operation along the time dimension is performed to reduce the number of nodes for easing optimization and reducing computational complexity. For clarity, we investigate different strategies for building the S-RFNs under single granularity setting (*i.e.*, SG-RAFA) and show the results in Table 3.

The spatial resolution of frame feature is $H \times W$, where *H=16, W=8* in this paper, resulting in 128 feature nodes for each frame feature F_t. The number of temporal frames is *T=8*. **Ours**: we obtain S-RFNs by fusing frame features F_t, $t = 1, \cdots, T$ along the time dimension and obtain a feature map with $16 \times 8 = 128$ feature nodes. **S-P**: we fuse feature nodes to obtain the reference set via average pooling along spatial dimensions. **T-P**: we perform average pooling along the temporal dimension with different ratios to obtain different settings in Table 3. **ST(**$16 \times 8 \times 8$**)**: we take all the spatial and temporal nodes as the reference set.

We have the following observations. (1) *Ours* outperforms schemes *S-P(8×1×8)*, *S-P(8×2×8)*, S-P(4×4×8), *S-P(8 × 4 × 8)* with spatial pooling by 1.2%, 1.2%, 1.0% and 0.9% in mAP respectively, where spatial pooling may remove too much useful information and result in inferior S-RFNs. (2) *Ours* also outperforms those with partial temporal pooling. The performance increases as the temporal pooling degree increases. (3) *Ours* outperforms the scheme *ST(*$16 \times 8 \times 8$*)* without any pooling by 0.8% in mAP. Using all nodes as reference results in a larger optimization space

Table 4. Comparison with non-local related schemes.

Models	Mars				
	mAP	R-1	R-5	R-10	R-20
NL(S)	83.2	86.6	95.9	97.1	97.9
NL(ST)	82.7	86.0	95.4	96.7	97.4
SG-RAFA	**85.1**	**87.8**	**96.1**	**97.8**	**98.5**

Table 5. Evaluation of the multi-granularity (MG) design when other attention methods are used on the extracted feature maps $F_t, t = 1, \cdots, T$. Granularity is set to $N = 4$.

Models	Mars				
	mAP	R-1	R-5	R-10	R-20
Baseline	82.1	85.9	95.1	96.5	97.3
RGA-SC	83.5	87.2	95.3	97.1	98.2
RGA-SC (MG)	**85.0**	**88.1**	**96.9**	**97.7**	**98.5**
SE	82.9	86.5	95.5	97.1	98.1
SE (MG)	**84.3**	**87.6**	95.4	97.1	98.1
CBAM	82.9	86.8	95.7	97.2	98.1
CBAM (MG)	**84.6**	**88.0**	95.7	97.2	97.9
MG-RAFA (Ours)	**85.9**	**88.8**	**97.0**	97.7	**98.5**

and the diversity of temporal patterns is complex which makes it difficult to learn. In contrast, through temporal average pooling, we reduce the pattern complexity and thus ease the learning difficulty and computational complexity.

Complexity. Thanks to the selection of S-RFNs, in comparison with *ST(*$16 \times 8 \times 8$*)* which uses all the feature nodes as reference, the computational complexity in terms of FLOPs for our aggregation module is reduced from 6.697G to 2.301G while the performance is 0.8% higher in mAP.

4.3.3 Comparison with Non-local

Non-local block [32] explores long-range context, which weighted sums the features from all positions to refine the current position feature. Both our approach and non-local can explore the global context. However, non-local block uses a *deterministic* way, *i.e.*, weighted summation (without parameters), to exploit the global information, which limits its capability. In contrast, ours could mine the structural pattern and semantics from the stacked relations by leveraging a learned model/function to inference the importance of a feature node as attention, being more flexible and having large optimization space.

Table 4 shows the comparisons to non-local schemes. We added a non-local module on the feature maps $F_t, t = 1, \cdots, T$ for feature refinement followed by spatio-temporal average pooling. *NL(ST)* denotes that non-local is performed on all the spatio-temporal features and *NL(S)* denotes that non-local is performed within each frame. Our *SG-RAFA* outperforms *NL(ST)* and *NL(S)* significantly by 2.4% and 1.9% in mAP, respectively. *NL(ST)* is inferior to *NL(S)* which may be caused by the optimization difficulty when spatio-temporal dimensions are jointly considered.

Table 6. Performance (%) comparison of our scheme with the state-of-the-art methods on three benchmark datasets[2].

Models	Mars				iLIDS-VID			PRID2011		
	mAP	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20	Rank-1	Rank-5	Rank-20
AMOC (TCSVT17)[19]	52.9	68.3	81.4	90.6	68.7	94.3	99.3	83.7	98.3	**100**
TriNet (ArXiv17)[8]	67.7	79.8	91.4	-	-	-	-	-	-	-
3D-Conv+NL (ACCV18)[17]	77.0	84.3	-	-	81.3	-	-	91.2	-	-
Snippt (CVPR18)[4]	76.1	86.3	94.7	98.2	85.4	96.7	99.5	93.0	99.3	**100**
DRSA (CVPR18)[15]	65.8	82.3	-	-	80.2	-	-	93.2	-	-
DuATM (CVPR18)[25]	62.3	78.7	-	-	-	-	-	-	-	-
M3D (AAAI19)[14]	74.1	84.4	93.8	97.7	74.0	94.3	-	94.4	**100**	-
STA (AAAI19)[5]	80.8	86.3	95.7	-	-	-	-	-	-	-
Attribute (CVPR19)[43]	78.2	87.0	95.4	**98.7**	86.3	97.4	**99.7**	93.9	99.5	**100**
GLTR (ICCV19)[13]	78.5	87.0	95.8	98.2	86.0	**98.0**	-	95.5	**100**	-
MG-RAFA (Ours)	**85.9**	**88.8**	**97.0**	98.5	**88.6**	98.0	**99.7**	**95.9**	99.7	**100**

4.3.4 Extension of MG Design to Other Attention

Different semantics could be suitably captured at different granularities (as illustrated in Figure 1). Our proposed multi-granularity design can also be applied to other attention mechanisms. We conduct experiments by applying several different attention designs on the extracted feature maps $F_t, t = 1, \cdots, T$ and show the results in Table 5. Compared with the single gruanularity versions, multi-granularity design brings 1.5%, 1.4%, and 1.7% improvements in mAP respectively for RGA-SC [40], SE [10], and CBAM [34], demonstrating the effectiveness of the proposed multi-granularity design. In addition, our proposed MG-RAFA outperforms *RGA-SC(MG)*, *SE(MG)*, *CBAM(MG)* by 0.9%, 1.6%, and 1.3% in mAP respectively.
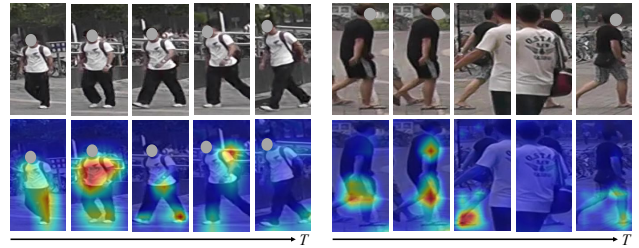
4.4. Comparison with State-of-the-arts

Table 6 shows that our *MG-RAFA* significantly outperforms the state-of-the-arts. On Mars, compared to STA [5], our method achieves **5.1%** improvements in mAP. On iLIDS-VID and PRID2011, ours outperforms the second best approach by 2.3% and 0.4% in Rank-1, respectively.

4.5. Visualization Analysis

We visualize the learned attention values at different spatial-temporal positions at different granularities in Figure 3. We have two observations from (a). (1) The learned attention tends to focus on different semantic regions from different frames, which gets rid of a lot of repetitions (redundancy). (2) Interestingly but not surprisingly, our attention is able to select the better represented areas and exclude the inferences (*e.g.*, see the 3rd and 4th columns of the right sub-figures in (a) where there are occlusions). We believe our attention modeling is an effective method to capture and learn discriminative spatial and temporal representation. (b) shows MG-RAFA captures different semantics at different granularities, which tends to capture more details at finer granularities and larger body parts at coarser granularities.

(a) Visualization on different frames at the 2nd granularity.

(b) Visualization of different granularities at a given time.

Figure 3. Visualization of our attention (a) across different frames, and (b) at different granularities. "G-1st" to "G-4th" denote the $1st$ to the $4th$ granularities, with their corresponding spatial resolutions of the attention masks for each frame as 16×8, 8×4, 4×2, 2×1, respectively. Here, we rescale the attention map of different spatial resolutions to the same spatial resolution for visualization.

5. Conclusion

In this paper, we propose a Multi-Granularity Reference-aided Attentive Feature Aggregation scheme (MG-RAFA) for video-based person re-identification, which effectively enhances discriminative features and suppresses identity-irrelevant features on the spatial and temporal feature representations. Particularly, to reduce the optimization difficulty, we propose to use a representative set of reference feature nodes (S-RFNs) for modeling the global relations. Moreover, we propose multi-gruanularity attention by exploring the relations at different granularity levels to capture semantics at different levels. Our scheme achieves the state-of-the-art performance on three benchmark datasets.

Acknowledgements

---

[2]We do not include results on DukeMTMC-VideoReID [45] since this dataset is not publicly released anymore.

References

[1] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification. *arXiv preprint arXiv:1801.05339*, 2018. 6

[2] Aaron F Bobick and James W Davis. The recognition of human movement using temporal templates. *TPAMI*, (3):257–267, 2001. 4

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2

[4] 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*, pages 1169–1178, 2018. 3, 6, 8

[5] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. AAAI, 2019. 1, 2, 3, 6, 8

[6] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. In *BMVC*, 2018. 1, 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 6, 8

[9] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 5

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 8

[11] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–231, 2012. 2

[12] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 1, 2

[13] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, pages 3958–3967, 2019. 8

[14] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, 2019. 2, 8

[15] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. 1, 2, 6, 8

[16] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 1, 2

[17] Xingyu Liao, Lingxiao He, Zhouwang Yang, and Chi Zhang. Video-based person re-identification via 3d convolutional networks and non-local attention. In *ACCV*, pages 620–634, 2018. 2, 8

[18] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, pages 3492–3506. 1, 2

[19] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *TCSVT*, 28(10):2788–2802, 2018. 2, 6, 8

[20] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, pages 5790–5799, 2017. 1, 3

[21] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019. 1

[22] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019. 1

[23] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bags of tricks and a strong baseline for deep person re-identification. *arXiv preprint arXiv:1903.07071*, 2019. 6

[24] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. 1, 2

[25] 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. 3, 8

[26] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 2

[27] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 2

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. 2018. 6

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5

[30] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018. 1, 2, 6

[31] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. Springer, 2014. 5

[32] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 7

[33] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018. 6

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 8

[35] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018. 2

[36] 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*, pages 4733–4742, 2017. 1, 3

[37] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716, 2016. 2

[38] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. *Pattern Recognition*, 86:143–155, 2019. 2

[39] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 6

[40] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention. *arXiv preprint arXiv:1904.02998*, 2019. 8

[41] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, 2020. 2, 3

[42] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3239–3248, 2017. 1, 2

[43] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR*, pages 4913–4922, 2019. 1, 3, 8

[44] 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*, pages 868–884. Springer, 2016. 5, 6

[45] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017. 8

[46] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 6

[47] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 4747–4756, 2017. 1, 3