# Watching You: Global-guided Reciprocal Learning for Video-based Person Re-identification

Xuehu Liu[1,2], Pingping Zhang[1,3,*], Chenyang Yu[1,2], Huchuan Lu[1,2,4,*], Xiaoyun Yang[5]

[1]School of Information and Communication Engineering, Dalian University of Technology
[2]Ningbo Institute, Dalian University of Technology
[3]School of Artificial Intelligence, Dalian University of Technology
[4]Pengcheng Lab, [5]Remark Holdings

{snowtiger, yuchenyang}@mail.dlut.edu.cn, {zhpp, lhchuan}@dlut.edu.cn, xyang@remarkholdings.com

## Abstract

*Video-based person re-identification (Re-ID) aims to automatically retrieve video sequences of the same person under non-overlapping cameras. To achieve this goal, it is the key to fully utilize abundant spatial and temporal cues in videos. Existing methods usually focus on the most conspicuous image regions, thus they may easily miss out fine-grained clues due to the person varieties in image sequences. To address above issues, in this paper, we propose a novel **Global-guided Reciprocal Learning (GRL)** framework for video-based person Re-ID. Specifically, we first propose a **Global-guided Correlation Estimation (GCE)** to generate feature correlation maps of local features and global features, which help to localize the high- and low-correlation regions for identifying the same person. After that, the discriminative features are disentangled into high-correlation features and low-correlation features under the guidance of the global representations. Moreover, a novel **Temporal Reciprocal Learning (TRL)** mechanism is designed to sequentially enhance the high-correlation semantic information and accumulate the low-correlation sub-critical clues. Extensive experiments are conducted on three public benchmarks. The experimental results indicate that our approach can achieve better performance than other state-of-the-art approaches. The code is released at https://github.com/flysnowtiger/GRL.*

## 1. Introduction

Person re-identification (Re-ID) aims to retrieve specific pedestrians cross different cameras at different times and places. Recently, this task has become a hot research topic due to its importance in advanced applications, such as safe

*Corresponding Authors

community, intelligent surveillance and criminal investigation. Compared with other related Re-ID tasks, video-based person Re-ID provides a video as the input to retrieve rather than a single image. Although videos can provide comprehensive appearance information, motion cues, pose variations in temporal, at the same time, they bring more illumination changes, complicated backgrounds and person occlusions in a clip. Thus, there are still many challenges for researches to handle in video-based person Re-ID.

Previous methods [30, 28, 17] can be coarsely summarized into two steps: spatial feature extraction and temporal feature aggregation. First, Convolutional Neural Networks (CNNs) are utilized to extract frame-level spatial features from each single image. Then, frame-level spatial features are temporally aggregated into a feature vector as the video representation to compute the similarity scores. Naturally, how to fully explore the discriminative spatial-temporal cues from multiple frames is seen as the key to tackle video-based person Re-ID. Generally speaking, the average pooling for spatial-temporal features can directly focus on main targets, but it has some obvious drawbacks, such as the inability to tackle the misalignment in temporal, the pollution of background noises, and the difficulty of capturing small but meaningful subjects in videos. To address these drawbacks, in recent years, researchers have proposed some rigid-partition-based methods or soft-attention-based methods to instead the direct average operation. These methods are beneficial to learn more discriminative and diverse local features, resulting in higher performance of video-based person Re-ID. However, previous methods generally ignore the role of the global features in whole person recognition while strengthening the local features. Based on this consideration, Zhang *et al.* [33] utilize the local affinities with respect to inference global features to help assign different weights to local features. Although effective, it tends to ignore inconspicuous yet fine-grained clues. Different

from [33], we correlate the global feature with the pixel-level local features in a frame to generate two correlation maps, which are utilized to disentangle generic features into high- and low-correlation features. Intuitively, features with high correlation mean they appear frequently in temporal and are spatially conspicuous. Features with low correlation mean they are inconspicuous and discontinuous yet meaningful. We further explore suitable strategies for disentangled features in temporal and fully mine fine-grained cues.

Based on above considerations, we proposed a novel Global-guided Reciprocal Learning (GRL) framework for video-based person Re-ID. The whole framework mainly consists of two key modules. To begin with, we proposed a Global-guided Correlation Estimation (GCE) module to estimate the correlation values of frame-level local features under the global guidance. With GCE, each frame-level feature map will be disentangled into two kinds of discriminative features with distinct correlation degrees. The one with high correlation, usually covers the most conspicuous and continuous visual information. Another with inverse correlation, as the supplement, is exploited to mine the fine-grained and sub-critical cues. Besides, we propose a novel Temporal Reciprocal Learning (TRL) module to fully exploit all the discriminative features in the forward and backward process. More specifically, for high-correlation features, we adopt a semantic enhancement strategy to mine spatial conspicuous and temporal aligned information. For low-correlation features, we introduce a temporal memory strategy to accumulate the discontinuous but discriminative cues frame by frame. In this way, our proposed method can not only explore the most conspicuous information from the high-correlation regions in a sequence, but also capture the sub-critical information from the low-correlation regions. Extensive experiments on public benchmarks demonstrate that our framework delivers better results than other state-of-the-art approaches.

In summary, our contributions are four folds:

- We propose a novel Global-guided Reciprocal Learning (GRL) framework for video-based person Re-ID.

- We propose a Global-guided Correlation Estimation module to generate the correlation maps under the guidance of video representations for disentanglement.

- We introduce a Temporal Reciprocal Learning (TRL) module to effectively capture the conspicuous information and the fine-grained clues in videos.

- Extensive experiments on public benchmarks demonstrate that our framework synthetically attains a better performance than several state-of-the-art methods.

## 2. Related Works

### 2.1. Video-based Person Re-identification

In recent years, with the rise of deep learning [6, 13], person Re-ID has gained a great success and the performance has been improved significantly. At the early stage of person Re-ID, researchers pay more attention to image-based person Re-ID. Recently, video-based person Re-ID is seen as a generalization of image-based person Re-ID task, and has drawn more and more researchers' interests. Generally, videos contain richer spatial and temporal information than still images. Thus, on the one hand, some existing methods [16, 22, 23, 7] concentrate on extracting attentive spatial features. On the other hand, some works [26, 20, 15] attempt to capture temporal information to strength the video representations. For example, Li et al. [16] employ a diverse set of spatial attention modules to consistently extract similar local patches across multiple images. Fu et al. [7] design an attention module to weight horizontal parts using a spatial-temporal map for more robust clip-level feature representations . Zhao et al. [34] propose a attribute-driven method for feature disentangling to learn various attribute-aware features. Liu et al. [18] propose a soft-parsing attention network and joint utilize a spatial pyramid non-local block to learn multiple semantic-aware aligned video representations. Zhang et al. [33] utilize a representative set of reference feature nodes for modeling the global relations and capturing the multi-granularity level semantics. In this paper, we attempt to estimate the correlation values of spatial features guided by the whole video representation, which is beneficial to cover the conspicuous visual cues in each frame. Besides, a novel temporal reciprocal learning mechanism is proposed to explore more discriminative information for video-based person Re-ID.

### 2.2. Temporal Feature Learning

For video-related tasks, such as video-based person Re-ID, action recognition, video segmentation [21] and so on, the temporal feature learning is seen as the core module in most algorithms. Typically, the temporal information modeling methods encode temporal relations or utilize temporal cues for video representation learning. Most of existing video-based Re-ID methods exploit optical flow [4, 2], Recurrent Neural Networks (RNN) [20, 10], or temporal pooling [35] for temporal feature learning. For the action recognition, Weng et al. [27] introduce a progressive enhancement module to sequentially excite the discriminative channels of frames. For video-based person Re-ID task, Mclaughlin et al. [20] introduce a recurrent architecture to pass the feature message of each frame for aggregating temporal information. Xu et al. [30] propose a joint spatial CNN and temporal RNN model for video-based person Re-ID. Zhang et al. [32] introduce a reinforcement learn-
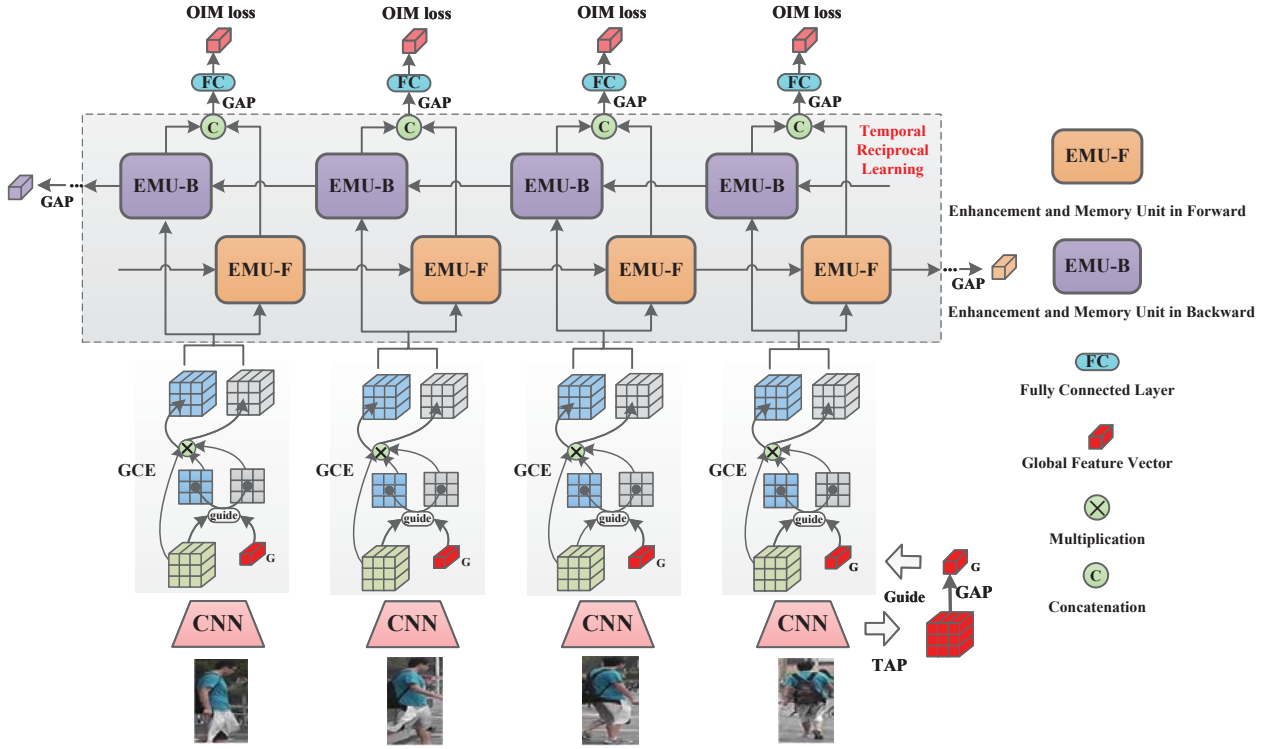
Figure 1. The overall structure of our proposed method. Given an image sequence, we firstly utilize ResNet-50 to extract frame-level feature maps. Then, frame-level features are aggregated by TAP and GAP to generate a video-level feature. With the guidance of video-level features, Global Correlation Estimation (GCE) is utilized to generate the correlation maps for disentanglement. Afterwards, the Temporal Reciprocating Learning (TRL) is introduced to enhance and accumulate disentangled features in forward and backward directions.

ing method for pairwise decision making. Dai *et al.* [5] design a temporal residual learning module to simultaneously extract the generic and specific features from consecutive frames. Liu *et al.* [19] design a refining recurrent unit and spatial-temporal integration module to integrate abundant spatial-temporal information. Compared with existing methods, our method adopts temporal reciprocal learning for bi-directional semantic feature enhancement and temporal information accumulation. Thus, the global-guided spatial features could focus on complementary objects, such as moving human body and key accessories.

## 3. Proposed Method

In this section, we introduce the proposed Global-guided Reciprocal Learning (GRL) framework. We first give an overview of the proposed GRL. Then, we elaborate the key modules in the following subsections.

### 3.1. Overview

The overall architecture of our proposed GRL is shown in Fig. 1. Our approach consists of frame-level feature extraction, global-guided feature disentanglement, temporal reciprocal learning. Given a video, we first use the Re-

stricted Random Sampling (RRS) [16] to generate training image frames. Then, we extract frame-level features by a pre-trained backbone network (ResNet-50 [8] in our work). After that, we adopt a Temporal Average Pooling (TAP) and a Global Average Pooling (GAP) to generate a video-level representation. With the guidance of the video-level representation, we design a Global-guided Correlation Estimation (GCE) to generate the correlation maps and disentangle the frame-level features to high- and low-correlation features. Afterwards, the Temporal Reciprocating Learning (TRL) is introduced to enhance and accumulate disentangled features in forward and backward directions. Finally, we introduce the Online Instance Matching (OIM) [29] loss and verification loss to optimize the whole network. By the GRL, our method can not only capture the conspicuous information but also mine meaningful fine-grained cues in sequences. In the test stage, the attentive pooled feature from the enhanced high-correlation vectors at different time steps and the accumulated low-correlation feature at the last time step are concatenated for the retrieval list.

### 3.2. Global-guided Feature Disentanglement

The attention mechanism has been widely adopted to tackle the misalignment in video-based person Re-ID. How-
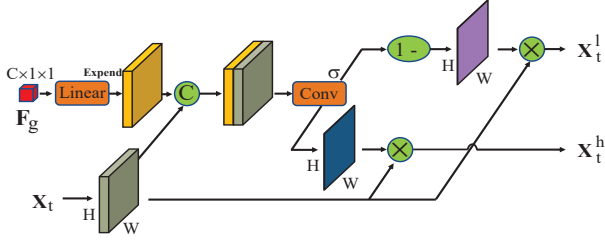
Figure 2. The proposed GCE module.

ever, existing attention-based methods lack global perceptions of the whole video and easily miss out fine-grained clues. To relieve this issue, we propose the GCE module to disentangle the spatial features into two complementary features. One of them highlights the conspicuous information in each frame, another keeps the fine-grained and sub-critical cues. Fig. 2 shows the structure of the proposed GCE module. Formally, given a video, we firstly sample $T$ frames $\nu = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_T\}$ as the inputs of our network. The ResNet-50 is utilized as the feature extractor to obtain a set of frame-level features $\chi = \{\mathbf{X}_t | t = 1, 2, ..., T\}$, where $\mathbf{X}_t \in R^{C \times H \times W}$, $H, W, C$ represents the height, width and the number of channels, respectively. Then, we utilize a TAP and a GAP to obtain the video-level representation

$$\mathbf{F}^g = \frac{\sum_{t,h,w=0}^{T,H,W} X_{t,h,w}}{H \times W \times T} \qquad (1)$$

$\mathbf{F}^g \in R^{C \times 1 \times 1}$ can coarsely represent the whole video.

To obtain global information, the proposed GCE takes both the frame-level features $\mathbf{X}_t$ and the video-level feature vector $\mathbf{F}^g$ as inputs. To guide the feature learning, the $\mathbf{F}^g$ is adhered with a linear projection and expanded to $\widetilde{\mathbf{F}}^g$, which has the same sizes to $\mathbf{X}_t$. The expanded features are concatenated with $\mathbf{X}_t$. Then, we integrate the global and local features, and jointly infer the degree of correlations. The correlation map $\mathbf{R}_t \in R^{1 \times H \times W}$ related to $\mathbf{X}_t$ under the global guidance can be computed by

$$\mathbf{R}_t = \sigma(\mathbf{W}_r([\widetilde{\mathbf{F}}^g, \mathbf{X}_t])) \qquad (2)$$

where $[\cdot, \cdot]$ represents the concatenation operation. $\mathbf{W}_r$ is learnable weight of two $1 \times 1$ convolutional layers inserted by Batch Normalization (BN) and ReLU activation. $\sigma$ represents the sigmoid activation function. By reversing the obtained correlation map, we can obtained a low-correlation map. Then, the correlation maps are multiplied with original frame-level features $\mathbf{X}_t$ to activate distinct local regions. Finally, under the guidance of global representation, we disentangle frame-level features into the high-correlation features $\mathbf{X}_t^h$ and the low-correlation features $\mathbf{X}_t^l$ by

$$\mathbf{X}_t^h = \mathbf{X}_t \odot \mathbf{R}_t \qquad (3)$$

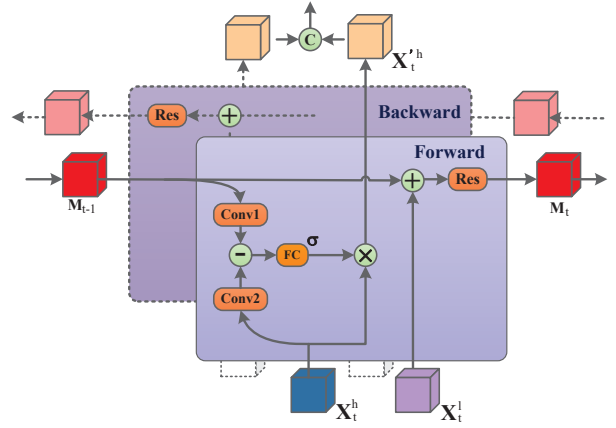$$\mathbf{X}_t^l = \mathbf{X}_t \odot (1 - \mathbf{R}_t) \qquad (4)$$



Figure 3. Enhancement and Memory Unit.

where $\odot$ represents element-wise multiplication, and $\mathbf{X}_t^h$, $\mathbf{X}_t^l \in R^{C \times H \times W}$. Based on above procedures, we disentangle generic features into two distinct features, which are different from previous methods on local feature extraction.

### 3.3. Temporal Reciprocal Learning

The temporal feature aggregation plays an important role in video-based person Re-ID. The GCE can highlight the informative regions in a global view. However, discontinuous fine-grained cues are easily missed out due to the visual varieties in a long sequence. To address this problem, we propose a novel Temporal Reciprocal Learning (TRL) mechanism to fully explore the discriminative information from the disentangled high- and low-correlation features. Considering the frame orders in videos, our TRL is designed for both forward and backward directions. More specifically, we introduce Enhancement and Memory Units (EMUs) to enhance high-correlation features and accumulate low-correlation features. Finally, the features passed through the forward and backward directions are integrated as the outputs of our TRL.

**Enhancement and Memory Unit.** As illustrated in Fig. 3, at the time step $t$, the EMU takes three inputs: the high-correlation features $\mathbf{X}_t^h$ and the low-correlation features $\mathbf{X}_t^l$, and the accumulated features $\mathbf{M}_{t-1}$ from previous time steps. In the enhancement block, we perform subtraction between the high-correlation features $\mathbf{X}_t^h$ and the accumulated features $\mathbf{M}_{t-1}$ to model the difference in semantics. Mathematically, the difference operation is defined as

$$\mathbf{D}_t = (f_2(\mathbf{M}_{t-1}) - f_1(\mathbf{X}_t^h))^2 \qquad (5)$$

where $f_1$ and $f_2$ represent two individual $1 \times 1$ convolution operations with ReLU activation, respectively. Then, the difference maps $\mathbf{D}_t$ are aggregated by GAP to generate an overall response for each channel, i.e., $\mathbf{d}_t \in R^C$. We introduce the channel attention for the feature selection, as

$$\mathbf{a}_t = \sigma(\mathbf{W}_c(\mathbf{d}_t)) \qquad (6)$$

$$\mathbf{X}_t^{'h} = (1 + \mathbf{a}_t) \odot \mathbf{X}_t^h \tag{7}$$

where $\mathbf{W}_c$ are the parameters for generating the channel weights. To fully exploit low-correlation features, we design a memory block to accumulate the fine-grained cues. Specifically, we first add low-correlation features $\mathbf{X}_t^l$ at $t$-th frame to the accumulated features $\mathbf{M}_{t-1}$ at $t-1$ step. Then, a residual block [8] is utilized for the next EMU.

$$\mathbf{M}_t = Res(\mathbf{M}_{t-1} + \mathbf{X}_t^l) \tag{8}$$

where $Res$ is the residual block in [8]. In the first time step, $\mathbf{M}_0$ is initialized with the mean of $\{\mathbf{X}_t^l\}_{t=1}^T$.

**Bi-directional Information Integration.** In TRL, we design a bi-directional learning mechanism for assembling more robust representations. The forward and backward directions are relative. For better understanding, we define the forward direction as the arranged order of video frames. The backward direction is opposite to forward direction. With the outputs of EMUs in forward and backward directions, we integrate them as the final video-level representation. Specifically, the enhanced features $\mathbf{F}_t^{h,1}$, $\mathbf{F}_t^{h,2}$, and the accumulated features $\mathbf{M}_T^1$ $\mathbf{M}_T^2$ in forward and backward are concatenated after GAP. Then, a fully connected layer is utilized to integrate the concatenated robust representations,

$$\mathbf{F}_t^h = \mathbf{W}_h([\mathbf{F}_t^{h,1}, \mathbf{F}_t^{h,2}]) \tag{9}$$

$$\mathbf{F}_T^l = \mathbf{W}_l([\mathbf{M}_T^1, \mathbf{M}_T^2]). \tag{10}$$

With the proposed temporal reciprocal learning mechanism, our method is able to progressively enhance the conspicuous features from high-correlation regions and adaptively mine the sub-critical details from low-correlation regions.

### 3.4. Training Schemes

In our work, we adopt a binary cross entropy loss and the Online Instance Matching loss (OIM) [29] to train the whole network following [2]. For each probe-gallery video vector pair $\{\mathbf{p}_j, \mathbf{g}_k\}$ in the training mini-batch, a binary cross entropy loss function can be utilized as

$$\mathcal{L}_{veri} = -\frac{1}{J} \sum_{n=1}^{J} y_j log([\mathbf{p}_j, \mathbf{g}_k]) + (1-y_j)log(1-[\mathbf{p}_j, \mathbf{g}_k]) \tag{11}$$

where $J$ is the number of sampled sequence pairs, $[\cdot, \cdot]$ denotes the similarity estimation function and $[\mathbf{p}_j, \mathbf{g}_k] \in (0, 1)$. $y_j$ denotes the ground-truth label of $\mathbf{p}_j$ and $\mathbf{g}_k$. Note that $y_j = 1$ if sequence $\mathbf{p}_j$ and $\mathbf{g}_k$ belong to the same person, otherwise $y_j = 0$.

Meanwhile, in our work, we use a multi-level training objective to deeply supervise our proposed modules, which consists of the frame-level OIM loss and video-level OIM loss. Instead of the conventional cross-entropy with a multi-class softmax layer, the OIM loss function uses a lookup

table to store features of all identifies in the training set. To learn informative and continuous features from different frames, in the temporal reciprocal learning, the features $\{\mathbf{F}_t^h\}_{t=1}^T$ enhanced by the enhancement block at $t$-th time step, are supervised by a frame-level OIM loss. The frame-level OIM loss can be defined as:

$$\mathcal{L}_f = -\frac{1}{N \times T} \sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{i=1}^{I} y_{t,n}^i log\left(\frac{e^{\mathbf{W}_i \mathbf{F}_{t,n}^h}}{\sum_{j=1}^{I} e^{\mathbf{W}_j \mathbf{F}_{t,n}^h}}\right) \tag{12}$$

where $\mathbf{F}_{t,n}^h$ indicates the enhanced high-correlation feature vector of the $t$-th image in $n$-th video. If the $t$-th image in $n$-th video belongs to the $i$-th person, $y_{t,n}^i = 1$, otherwise $y_{t,n}^i = 0$. $\mathbf{W}_i$ are the coefficients associated with the feature embedding of the $i$-th person, which are online updated with the frame-wise feature vectors of the $i$-th person. Meanwhile, the feature $\mathbf{F}_T^l$ accumulated by the memory block at last time step, is supervised by video-level OIM loss, which attempts to progressively collect all the sub-critical details from the low-correlation regions.

$$\mathcal{L}_v = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I} y_n^i log\left(\frac{e^{\mathbf{W}_i \mathbf{F}_T^l}}{\sum_{j=1}^{I} e^{\mathbf{W}_j \mathbf{F}_T^l}}\right) \tag{13}$$

The total loss is a combination of the frame-level OIM loss, the video-level OIM loss and the verification loss.

$$\mathcal{L} = \lambda_1 \mathcal{L}_f + \lambda_2 \mathcal{L}_v + \lambda_3 \mathcal{L}_{veri} \tag{14}$$

## 4. Experiments

### 4.1. Datasets and Evaluation Protocols

To evaluate the performance of our proposed method, we adopt three widely-used benchmarks, *i.e.*, iLIDS-VID [26], PRID-2011 [9] and MARS [35]. iLIDS-VID [26] dataset is a small dataset, which consists of 600 video sequences of 300 different identities. Two cameras are used to collect images. Each video sequence contains 23 to 192 frames. PRID-2011 [9] dataset consists of 400 image sequences for 200 identities from two non-overlapping cameras. The sequence lengths range from 5 to 675 frames, with an average of 100. Following previous practice [26], we only utilize the sequence pairs with more than 21 frames. MARS [35] is one of large-scale datasets, and consists of 1,261 identities around 18,000 video sequences. All the video sequences are captured by at least 2 cameras.

For evaluation, we follow previous works and adopt the Cumulative Matching Characteristic (CMC) table and mean Average Precision (mAP) to evaluate the performance. In terms of iLIDS-VID and PRID2011, we only report the cumulative re-identification accuracy because that there only contains a single correct match in the gallery set.

Table 1. Comparison with state-of-the-art video-based person Re-ID methods on MARS, iLIDS-VID and PRID2011.

| Methods | Source | MARS | | | | iLIDS-VID | | | PRID2011 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mAP | Rank-1 | Rank-5 | Rank-20 | Rank-1 | Rank-5 | Rank-20 | Rank-1 | Rank-5 | Rank-20 |
| SeeForest [36] | CVPR17 | 50.7 | 70.6 | 90.0 | 97.6 | 55.2 | 86.5 | 97.0 | 79.4 | 94.4 | 99.3 |
| ASTPN [30] | ICCV17 | - | 44 | 70 | 81 | 62 | 86 | 98 | 77 | 95 | 99 |
| Snippet [2] | CVPR18 | 76.1 | 86.3 | 94.7 | 98.2 | 85.4 | 96.7 | 99.5 | 93.0 | 99.3 | 100 |
| STAN [16] | CVPR18 | 65.8 | 82.3 | - | - | 80.2 | - | - | 93.2 | - | - |
| STMP [19] | AAAI19 | 72.7 | 84.4 | 93.2 | 96.3 | 84.3 | 96.8 | 99.5 | 92.7 | 98.8 | 99.8 |
| M3D [15] | AAAI19 | 74.0 | 84.3 | 93.8 | 97.7 | 74.0 | 94.3 | - | 94.4 | 100 | - |
| STA [7] | AAAI19 | 80.8 | 86.3 | 95.7 | 98.1 | - | - | - | - | - | - |
| Attribute [34] | CVPR19 | 78.2 | 87.0 | 95.4 | **98.7** | 86.3 | 87.4 | 99.7 | 93.9 | 99.5 | 100 |
| VRSTC [12] | CVPR19 | 82.3 | 88.5 | 96.5 | 97.4 | 83.4 | 95.5 | 99.5 | - | - | - |
| GLTR [14] | ICCV19 | 78.5 | 87.0 | 95.8 | 98.2 | 86.0 | 98.0 | - | 95.5 | **100** | - |
| COSAM [24] | ICCV19 | 79.9 | 84.9 | 95.5 | 97.9 | 79.6 | 95.3 | - | - | - | - |
| MGRA [33] | CVPR20 | **85.9** | 88.8 | **97.0** | <u>98.5</u> | <u>88.6</u> | <u>98.0</u> | <u>99.7</u> | <u>95.9</u> | 99.7 | <u>100</u> |
| STGCN [31] | CVPR20 | 83.7 | 89.9 | - | - | - | - | - | - | - | - |
| AFA [3] | ECCV20 | 82.9 | <u>90.2</u> | 96.6 | - | 88.5 | 96.8 | 99.7 | - | - | - |
| TCLNet [11] | ECCV20 | <u>85.1</u> | 89.8 | - | - | 86.6 | - | - | - | - | - |
| Ours | – | 84.8 | **91.0** | <u>96.7</u> | 98.4 | **90.4** | **98.3** | **99.8** | 96.2 | <u>99.7</u> | **100** |

## 4.2. Implementation Details

We implement our framework based on the Pytorch[1] toolbox. The experimental devices include an Intel i4790 CPU and two NVIDIA GTX 2080ti GPUs (12G memory). To generate training sequences, we employ the RRS strategy [7], and divide each video sequence into 8 chunks with equal duration. Experimentally, we set the batchsize = 16 and $T = 8$. Each image in a sequence is resized to $256 \times 128$ and the input sequences are augmented by random cropping, horizontal flipping and random erasing. To provide a number of positive and negative sequence pairs in each training mini-batch, we randomly sampled the half batchsize sequences firstly. For one sampled sequences, we select another sequence with the same identify but under different cameras to fill the total batch. In this way, there is at least one positive sample for any sequence in a minibatch. The ResNet-50 [8] pre-trained on the ImageNet dataset [6] is used as our backbone network. Following previous works [25], we remove the last spatial down-sampling operation to increase the feature resolution. During training, we train our network for 50 epochs combining with the multi-level OIM losses and a binary cross-entropy loss. The whole network is updated by stochastic gradient descent [1] algorithm with an initial learning rate of $10^{-3}$, weight decay of $5 \times 10^{-4}$ and nesterov momentum of 0.9. The learning rate is decayed by 10 at every 15 epochs.

## 4.3. Comparison with State-of-the-arts

In this section, we compare the proposed approach with other state-of-the-art methods on three video-based person Re-ID benchmarks. Experimental results are reported in

Tab. 1. On MARS dataset, the mAP and Rank-1 accuracy of our proposed method are $84.8\%$ and $91.0\%$, respectively. Besides, our method achieves $90.4\%$ and $96.2\%$ of the Rank-1 accuracy on iLIDS-VID dataset and PRID2011 dataset. The Rank-1 accuracy of our method outperforms all the compared methods, showing significant improvements over several state-of-the-art methods. We note that the MGRA [33] also employs the global view for the video-based person Re-ID task. It gains remarkable $85.9\%$ mAP on MARS dataset. Different from MGRA, our method utilizes the global representations to estimate two correlation maps for the feature disentanglement on spatial features. With the reciprocal learning, our method could fully take advantages of the disentangled features, and explore more informative and fine-grained cues via the high-correlation maps and low-correlation maps. Thereby, our method surpasses MGRA by $2.2\%$, $1.8\%$ and $0.3$ in terms of Rank-1 accuracy on MARS, iLIDS-VID and PRID2011, respectively. Meanwhile, it is worth noting that those methods, ASTPN [20], STMP [19] and GLTR [14], explore the temporal learning for video-based person Re-ID. ASTPN [20] utilizes a temporal RNN to model the temporal information for video representations. STMP [19] introduces a refining recurrent unit to recover the missing parts by referring historical frames. GLTR [14] employs dilated temporal convolutions to capture the multi-granular temporal dependencies and aggregates short and long-term temporal cues for global-local temporal representations. Compared with these methods, our proposed method achieves better results on three public datasets. More specifically, compared with GLTR [14], our method improves the performances by $6.3\%$ and $4.0\%$ in terms of mAP and Rank-1 accuracy on MARS dataset. In summary, compared with

---

[1]https://pytorch.org/

Table 2. Ablation results of key components on three benchmarks.

| Methods | Feat. to test | MARS | | iLIDS-VID | | PRID2011 | |
|---------|---------------|------|--------|-----------|--------|----------|--------|
| | | mAP | Rank-1 | Rank-1 | Rank-5 | Rank-1 | Rank-5 |
| Baseline | $F^g$ | 81.2 | 88.5 | 87.1 | 97.2 | 93.5 | 98.7 |
| + GCE | $F^l$ | 81.9 | 88.9 | 88.3 | 97.4 | 93.9 | 99.0 |
| | $F^h$ | 83.0 | 89.5 | 89.2 | 98.0 | 94.7 | 99.2 |
| | $F^l, F^h$ | 83.3 | 90.1 | 89.5 | 97.9 | 95.0 | 99.5 |
| + TRL | $F^l$ | 82.2 | 88.4 | 88.6 | 97.8 | 94.8 | 99.6 |
| | $F^h$ | 84.0 | 90.4 | **90.5** | 97.9 | 95.4 | 99.6 |
| | $F^h, F^l$ | **84.8** | **91.0** | 90.4 | **98.3** | **96.2** | **99.7** |

existing methods, our method utilizes the global information to guide the feature disentanglement. In additional, we adopt two strategies to mine richer cues for temporal learning, which can fully exploit the spatial-temporal information for more discriminative video representations. These experimental results validate the superiority of our method.

### 4.4. Ablation Study

In this subsection, we conduct experiments to verify the effectiveness of the proposed methods. All the models are trained and evaluated on MARS, iLIDS-VID and PRID2011 datasets. Results are shown in Tab. 1-5. In these tables, "Baseline" represents the backbone trained only with video-level OIM loss on the global branch, in which TAP and GAP are applied on the frame-level features.

**Effects of Key Components.** The ablation results of key components are reported in Tab. 2. In this table, $F^g$ denotes the global feature vector without disentanglements. $F^l$ denotes the final feature vector with disentangled low-correlation features, and is supervised with a video-level OIM loss. $F^h$ denotes the final feature vector with disentangled high-correlation features, and is supervised with a frame-level OIM loss. "+GCE" means that we add the global-guided correlation estimation to guide the disentanglement of spatial features. One can see that the performance has a significant improvement after disentanglement. The disentangled high-correlation features increase the Rank-1 accuracy by 1.6%, 2.4% and 1.5% on MARS, iLIDS-VID and PRID2011, respectively. Thus, it is beneficial to guide the feature disentanglement under a global view. "+TRL" means that the temporal reciprocal learning with bi-directions is used to enhance and accumulate temporal information. Compared with the "+GCE" model, our proposed TRL mechanism can further improve the mAP by 1.5% and the Rank-1 accuracy by 0.9% on MARS. As shown in Tab. 2, the combination of the low- and high-correlation features can further boost the performance. The above results clearly demonstrate the effectiveness of our proposed GCE and TRL modules.

**Effects of Enhancement and Memory Unit.** We also perform experiments to verify the effectiveness of EMU. The results on MARS are shown in Tab. 3. "GRL" means that our proposed GRL approach with complete EMUs. "-Memory Block" denotes that the memory blocks in EMUs

Table 3. Ablation results of EMU on MARS.

| Methods | mAP | Rank-1 | Rank-5 | Rank-20 |
|---------|-----|--------|--------|---------|
| Baseline | 81.2 | 88.5 | 95.5 | 97.9 |
| GRL | 84.8 | 91.0 | 96.7 | 98.4 |
| - Memory Block | 84.2 | 90.2 | 96.3 | 98.2 |
| - Enhancement Block | 83.4 | 90.1 | 96.5 | 98.3 |

Table 4. Ablation results of the sequence length.

| Methods | Length | mAP | Rank-1 | Rank-5 | Rank-20 |
|---------|--------|-----|--------|--------|---------|
| Baseline | 8 | 81.2 | 88.5 | 95.5 | 97.9 |
| GRL | 4 | 83.0 | 89.4 | 96.1 | 98.2 |
| | 6 | 83.5 | 90.1 | 96.7 | 98.4 |
| | 8 | 84.8 | 91.0 | 96.7 | 98.4 |
| | 10 | 83.9 | 90.2 | 96.5 | 98.4 |

Table 5. Ablation results of the direction order.

| Methods | Direction | mAP | Rank-1 | Rank-5 | Rank-20 |
|---------|-----------|-----|--------|--------|---------|
| Baseline | | 81.2 | 88.5 | 95.5 | 97.9 |
| GRL | Forward | 84.0 | 89.9 | 96.5 | 98.3 |
| | Backward | 83.7 | 90.0 | 96.5 | 98.3 |
| | Bi-direction | 84.8 | 91.0 | 96.7 | 98.4 |

are removed. The results show that, there are slight decreases in terms of mAP and Rank-1 accuracy on MARS. "-Enhancement Block" denotes that the enhancement blocks in EMUs are removed, in which the high-correlation features $\mathbf{X}_t^h$ are supervised by the frame-level OIM loss without channel attention. The mAP and Rank-1 accuracy drop with 1.4% and 0.9% on MARS. From the results, we can find that both our enhancement block and memory block are beneficial to learn more discriminative spatial features.

**Effects of Different Sequence Lengths.** We train and test our bi-directional global-guided reciprocal learning with various sequence lengths $T$. The results are shown in Tab. 4. From the results, we can see that increasing the length of sequence gains better performance and the length of 8 gets best performance. One possible reason is that our temporal reciprocating learning could capture more fine-grained cues with the increments of the sequence length. However, too long sequences are not good for training the temporal reciprocal learning module.

**Effects of Temporal Directions.** We perform additional experiments to verify the effectiveness of temporal directions in GRL. As shown in Tab. 5, the proposed temporal learning with forward or backward direction gains similar results. Besides, the bi-directional reciprocal learning shows higher performances, which benefits from the combination of forward and backward temporal learning. The effectiveness indicates that the aggregated features by reciprocating learning are more robust for identification.

**Effects of Multi-level OIM Losses.** The ablation results of multi-level OIM losses on MARS are reported in Tab. 6. The "F-OIM" denotes the frame-level OIM loss is deployed for each frame. The "V-OIM" denotes the video-level OIM
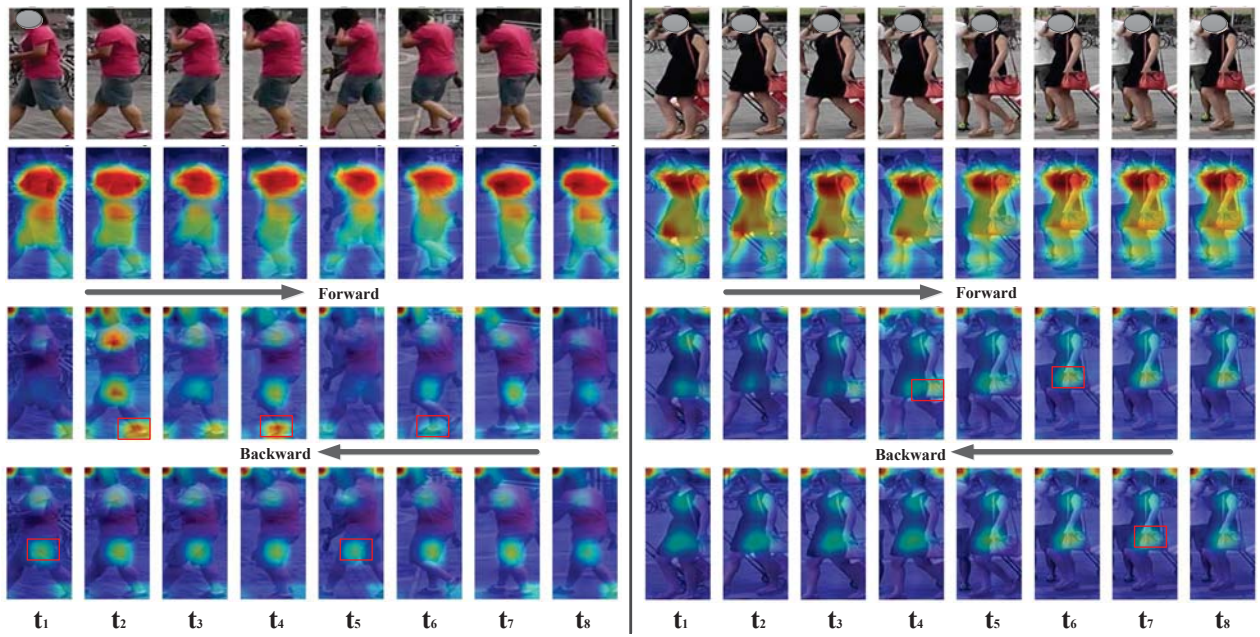
Figure 4. The visualization of the high-correlation maps and the accumulated low-correlation features at different time steps. The top images are raw images in video sequences. The heat maps in the second row are high-correlation maps $\{R_t\}_{t=1}^{T=8}$. The heat maps in the third and fourth rows are channel activation maps of accumulated features, $\{M_t^1\}_{t=1}^{T=8}$ and $\{M_t^2\}_{t=1}^{T=8}$ in the forward and backward process.

Table 6. Ablation results of multi-level OIM losses on MARS.

| Methods | Losses | mAP | Rank-1 | Rank-5 | Rank-20 |
|---------|--------|------|--------|--------|---------|
| GRL | V-OIM | 82.6 | 89.5 | 96.3 | 98.1 |
| | F-OIM | 83.5 | 90.1 | 96.4 | 98.3 |
| | V&F-OIM | 84.8 | 91.0 | 96.7 | 98.4 |

loss is utilized for each video. As shown in Tab. 6, higher performances are achieved when combining the frame-level and video-level OIM losses. It demonstrates that multi-level losses could better optimize our proposed GRL.

### 4.5. Visualization Analysis

We visualize the high-correlation maps and the accumulated low-correlation features in Fig. 4. Generally, features with high-correlations mean that they appear frequently in temporal and are spatially conspicuous. Features with low-correlations mean that they are inconspicuous and discontinuous yet meaningful. As shown in Fig. 4, the second row represents the high-correlation maps obtained from Equ.(2), covering the main and conspicuous regions, *e.g.*, human upper body. The third and forth rows show the accumulated low-correlation features, covering discontinuous but fine-grained cues, *e.g.*, the bags or shoes. Compared with the features learned from the high-correlation maps, the features from low-correlation maps in forward or backward process, could capture the incoherent and meaningful cues, such as shoes or bags, with red bounding boxes. Meanwhile, we can find that, at the same time step, there are

some variations among the features from the forward and backward process. Thus, it is useful to assemble more discriminative information. The visual maps further validate that our method could highlight the most conspicuous and aligned information in temporal and capture the sub-critical clues in spatial, simultaneously.

## 5. Conclusion

In this paper, we propose a novel global-guided reciprocal learning framework for video-based person Re-ID. We design a GCE module to estimate the correlation maps of spatial features under the global guidance. Then, spatial features are disentangled into the high- and low-correlation features. Besides, we propose a novel TRL module, in which multiple enhancement and memory units are designed for temporal learning. Based on the proposed modules, our approach could not only enhance the conspicuous information from the high-correlation regions, but also accumulative fine-grained cues from the low-correlation features. Extensive experiments on public benchmarks show that our framework outperforms several state-of-the-arts.

# References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186. 2010. 6

[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*, pages 1169–1178, 2018. 2, 5, 6

[3] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, pages 660–676, 2020. 6

[4] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *CVPR*, pages 1983–1991, 2017. 2

[5] Ju Dai, Pingping Zhang, Dong Wang, Huchuan Lu, and H Wang. Video person re-identification by temporal residual learning. *TIP*, 28:1366–1377, 2019. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2, 6

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

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 5, 6

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

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *NC*, pages 1735–1780, 1997. 2

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. *arXiv preprint arXiv:2007.09357*, 2020. 6

[12] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, pages 7183–7192, 2019. 6

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015. 2

[14] 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. 6

[15] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, pages 8618–8625, 2019. 2, 6

[16] 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. 2, 3, 6

[17] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, pages 3810–3818, 2015. 1

[18] Xuehu Liu, Pingping Zhang, Dai Ju, Huchuan Lu, and Yang Xiaoyun. Joint pyramid non-local and soft parsing attention for video-based person re-identification. *TIP*, 2020. 2

[19] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, pages 8786–8793, 2019. 3, 6

[20] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. 2, 6

[21] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *CVPR*, pages 10366–10375, 2020. 2

[22] 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*, pages 5363–5372, 2018. 2

[23] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. *arXiv preprint arXiv:1711.08766*, 2017. 2

[24] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, pages 562–572, 2019. 6

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 6

[26] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014. 2, 5

[27] Junwu Weng, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Xudong Jiang, and Junsong Yuan. Temporal distinct representation learning for action recognition. *arXiv preprint arXiv:2007.07626*, 2020. 2

[28] Lin Wu, Chunhua Shen, and Anton van den Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv:1606.01609*, 2016. 1

[29] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3415–3424, 2017. 3, 5

[30] 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, 2, 6

[31] 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*, pages 3289–3299, 2020. 6

[32] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *CVPR*, pages 6781–6789, 2018. 2

[33] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, pages 10407–10416, 2020. 1, 2, 6

[34] 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. 2, 6

[35] 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, 2016. 2, 5

[36] 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. 6