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

Xuehu Liu†§ Pingping Zhang†§ Chenyang Yu†§ Huchuan Lu†§* Xiaoyun Yang*
†School of Information and Communication Engineering, Dalian University of Technology,
§Ningbo Institute, Dalian University of Technology,
†School of Artificial Intelligence, Dalian University of Technology,
⋆China Science IntelliCloud Technology Co.Ltd

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 on three public benchmarks indicate that our approach can achieve better performance than other state-of-the-art approaches.

1. Introduction

Person re-identification (Re-ID) aims to retrieve specific pedestrian 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 community, intelligent surveillance and criminal investigation. In the past decade, the researches have made a great progress in this filed, and derive some interesting tasks associated with person Re-ID, such as image-based person Re-ID, video-based person Re-ID, person search and so on. 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, it brings 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. Technically, the video representation, obtained by Global Average Pooling (GAP) followed by Temporal Average Pooling (TAP), can directly focus on the important information of persons in the sequence. However, the average operation 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 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. For examples, Li et al. [16] do not consider the relationships between global and local features, and lacks global guidance for the local feature learning on each frame. Based on this consideration, Zhang et al. [33] utilize the local affinities with respect to

*Prof. Lu is the corresponding author. Email: lhchuan@dlut.edu.cn.
Inference global features to help assign different weights to local features. Different from the guidance for estimating weights of each node feature in [33], we correlate the video-level global features with the pixel-level local features in a frame to generate two correlation maps. Unlike the previous approaches using spatial self-attention mechanism, we utilize the global information to estimate the correlation degree in a frame. High- or low-correlation areas are adaptively mined according to the learned correlation maps. Intuitively, the higher the correlation values are, the more stable in temporal and more important in spatial the visual cues are. If the correlation values are low, the visual cues may be not continuous in temporal.

Based on above fact, 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 generate the correlation values of frame-level local features under the global guidance. With this new module, the frame-level feature maps 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. The other with inverse-correlation, as the supplement, is exploited to mine the 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 reinforcement strategy to mine spatial conspicuous and temporal aligned features. 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 highlighted regions in a sequence, but also capture the meaningful 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 new Global-guided Correlation Estimation module to generate the feature correlation degrees under the guidance of video-level representations.
- 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 most attention on 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 [26, 20, 15] attempt to capture temporal information to strength the video representations. On the other hand, some works [16, 22, 23, 7] concentrate on extracting attentive spatial features. 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 multi-granularity level attention to capture semantics at different levels. In this paper, we attempt to estimate the correlation maps 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 encodes the temporal relations or utilizes the temporal cues for video representation learning. Most of video-based existing 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. In video-based person Re-ID filed, 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. Dai et al. [5] design a temporal residual learning...
module to simultaneously extract the generic and specific features from consecutive frames. Zhang et al. [32] introduce a reinforcement learning method for pairwise decision making. Liu et al. [19] design a refining recurrent unit and spatial-temporal integration module integrate abundant spatial-temporal information. Compared with existing methods, our method adopts temporal reciprocal learning for bi-directional 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 Restricted Random Sampling (RRS) [16] to generate training image frames. Then, we extract frame-level features by a pre-trained backbone network (e.g., ResNet50 [8] in our work). Then, we adopt a Temporal Average Pooling (TAP) and a Global Average Pooling (GAP) to generate video-level representations based on the frame-level features. With the guidance of video-level representations, we design a Global-guided Correlation Estimation module to generate the correlation map for disentangling high- and low-correlation features according to the correlation degree. Afterwards, the Temporal Reciprocating Learning (TRL) is introduced to enhance and accumulate disentangled features in forward and backward directions. Finally, video-level and frame-level OIM losses are used to supervise the model training for Re-ID tasks.

Figure 1. The overall structure of our proposed method. Given an image sequence, we first extract frame-level feature maps by a pre-trained backbone network (e.g., ResNet50 [8]). Then, frame-level features are aggregated by TAP and GAP to generate video-level features. With the guidance of video-level features, Global Correlation Estimation (GCE) module is used to generate the correlation map for disentangling high- and low-correlation features according to the correlation degree. Afterwards, the Temporal Reciprocating Learning (TRL) is introduced to enhance and accumulate disentangled features in forward and backward directions. Finally, video-level and frame-level OIM losses are used to supervise the model training for Re-ID tasks.
However, computing attention scores from a single frame lacks global perceptions of the whole video. To relieve this issue, we propose the GCE module to highlight the conspicuous features in each frame and ignore the distractive information. Fig. 2 shows the structure of the proposed GCE module. Formally, given a video, we first sample T frames, \( \nu = \{ \mathbf{I}^1, \mathbf{I}^2, ..., \mathbf{I}^T \} \) as the input of our network. The ResNet-50 is used as the single frame feature extractor to obtain a set of frame-level features \( \mathbf{X}^t \). \( H, W, C \) represent the height, width and the number of channels, respectively.

Given the video-level representation, the proposed GCE takes the frame-level features \( \mathbf{X}^t \) as inputs. To guide the feature learning, the \( \mathbf{F}^g \) is first expanded to \( \mathbf{F}^g \) with the same shape of \( \mathbf{X}^t \). Then, we integrate the global and local features, and jointly infer the degree of correlation of them. The correlation map \( \mathbf{R}^t \) is computed by

\[
\mathbf{R}^t = \sigma(\mathbf{W}_r([\mathbf{F}^g, \mathbf{X}^t]))
\]

where \([\cdot, \cdot]\) represents the concatenation operation. \( \mathbf{W}_r \) is learnable weight of one \( 1 \times 1 \) convolution layer inserted by Batch Normalization (BN) and ReLU activation. \( \sigma \) represents the sigmoid activation function. By reversing the obtained correlation map, we can obtain a low-correlation map.

Finally, under the guidance, we disentangle frame-level features into the high-correlation features and the low-correlation features. The step-by-step disentanglement is illustrated in Fig. 3. At the time step \( t \), the high-correlation features \( \mathbf{X}^{h,t} \) and the low-correlation features \( \mathbf{X}^{l,t} \) are obtained. More specifically, we introduce the Enhancement and Memory Units (EMU) to enhance the high-correlation features and accumulate the low-correlation features.

\[
\mathbf{X}^{h,t} = \mathbf{X}^t \odot \mathbf{R}^t
\]

(3)

\[
\mathbf{X}^{l,t} = \mathbf{X}^t \odot (1 - \mathbf{R}^t)
\]

(4)

where \( \odot \) represents element-wise multiplication, and \( \mathbf{X}^{h,t}, \mathbf{X}^{l,t} \in \mathbb{R}^{C \times H \times W} \). Based on above procedures, we disentangle the frame-level features into two distinct features, which is 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 proposed GCE can highlight the informative regions under a global view. However, the sub-critical misaligned fine-grained cues are easily missed out due to the visual varieties in a long sequence. To address this problem, we proposed a novel Temporal Reciprocal Learning (TRL) mechanism to fully exploit the discriminative information from the high- and low-correlation features disentangled by GCE. Considering to the frame orders in videos, the TRL is designed for both forward and backward. More specifically, we introduce the Enhancement and Memory Units (EMU) to enhance the high-correlation features and accumulate the low-correlation features. Finally, the features passed through the forward and backward directions are integrated as the outputs of the TRL module.

**Enhancement and Memory Unit.** As illustrated in Fig. 3, at the time step \( t \), the EMU takes three inputs: the high-correlation features \( \mathbf{X}^{h,t} \) and the low-correlation features \( \mathbf{X}^{l,t} \), and the accumulated features \( \mathbf{M}^{t-1} \) from previous time steps. Then, we use the difference of feature maps to model the difference in semantics. Thus, the repeated semantic-aware feature between the high-correlation feature \( \mathbf{X}^{h,t} \) and the accumulative representation \( \mathbf{M}^{t-1} \), could be suppressed.
defined as
\[
D^t = f_2(M^{t-1}) - f_1(X^{h,t})
\]
where \(f_1\) and \(f_2\) represent two individual \(1 \times 1\) convolution operations with ReLU activation, respectively. Then, the difference maps \(D_t\) are aggregated by GAP to get overall response for each channel, i.e., \(d^t \in \mathbb{R}^{C_t}\). We introduce the channel attention for the feature selection, as
\[
a^t = \sigma(W_c(d^t)) \tag{6}
\]
\[
X^{h,t} = a^t \odot X^{h,t} \tag{7}
\]
\[
M^{t} = a^t \odot M^t \tag{8}
\]
where \(W_c\) are the parameters for generating the channel weights. \(M^t\) is the final enhancement features. To fully exploit low-correlation features, we design a memory block to accumulate the sub-critical cues. Specifically, we first concatenate the enhanced features \(M^{t-1}\) at \(t-1\) steps and the low-correlation features \(X^{l,t}\) at \(t\)-th frame. Then, a residual block [8] is utilized to integrate current low-correlation features and accumulated \(M_t\) to the next EMU.
\[
M' = Res(\{M^{t-1}, X^{l,t}\}) \tag{9}
\]
where \(Res\) is the residual block in [8]. In the first time step, \(M_0\) is initialized with the mean of \(\{X^{l,t}\}_{1=1}^{T}\).

**Bi-directional Information Integration.** In previous part, we design a bi-directional learning mechanism for assembling more robust representations. With the outputs of EMUs in forward and backward directions, we integrate them as the final video-level representation. Specially, the enhanced feature maps \(F^{h,t}_1, F^{h,t}_2\), the accumulated features \(M^{l}_1, M^{l}_2\) in forward and backward are concatenated after GAP. Then, a fully connected layer is utilized to integrate the concatenated robust representations.
\[
F^{h,t} = W_h([F^{h,t}_1, F^{h,t}_2]) \tag{10}
\]
\[
F^{l,T} = W_l([M^{l}_1, M^{l}_2]) \tag{11}
\]
With our proposed the temporal reciprocal learning mechanism, our method can 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 this paper, 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 sequence vector pair \(\{p_j, g_k\}\) in the training mini-batch, a binary cross entropy loss function can be utilized as
\[
\mathcal{L}_{veri} = -\frac{1}{J} \sum_{n=1}^{J} y_j \log([p_j, g_k]) + (1 - y_j) \log(1 - [p_j, g_k]) \tag{12}
\]
where \(J\) is the number of sampled sequence pairs, \([\cdot, \cdot]\) denotes the similarity estimation function and \([p_j, g_k] \in (0, 1)\). \(y_j\) denotes the ground-truth label of \(p_j\) and \(g_k\). Define \(y_j = 1\) if sequence \(p_j\) and \(g_k\) belong to the same person, otherwise \(y_j = 0\).

Meanwhile, in our work, we use 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 identities in the training set. In the temporal reciprocal learning, the feature \(F^{h,t}\) enhanced by the enhancement block at time index \(t\), is supervised by frame-level OIM loss, which is related to high-correlation map and aims to learn informative and continuous feature from different frames. The frame-level OIM loss can be defined as:
\[
\mathcal{L}_f = -\frac{1}{N \times T} \sum_{n=1}^{N} \sum_{l=1}^{T} \sum_{i=1}^{I} y_{n,i,t} \log\left(\frac{e^{W_n F^{h,t}_i}}{\sum_{j=1}^{N} e^{W_n F^{h,t}_j}}\right) \tag{13}
\]
where \(F^{h,t}_n\) indicates enhanced high-correlation feature vector of \(t\)-th image in \(n\)-th video. If the \(t\)-th image in \(n\)-th video belongs to the \(i\)-th person, \(y_{n,i,t} = 1\), otherwise \(y_{n,i,t} = 0\). \(W_n\) are the coefficients associated with the feature embedding of the \(i\)-th person, which is online updated with the the frame-wise feature vector of the \(i\)-th person. Meanwhile, the feature \(F^{l,T}\) accumulated by the memory block at the time step, is supervised by video-level OIM loss, which attempts to progressively collect all the sub-critical details from the low-correlation regions.
\[
\mathcal{L}_v = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{I} y_{n,i,t} \log\left(\frac{e^{W_n F^{l,T}_i}}{\sum_{j=1}^{J} e^{W_n F^{l,T}_j}}\right) \tag{14}
\]

The total loss for training is a combination of the frame-level OIM loss, the video-level OIM loss and the verification loss.
\[
\mathcal{L} = \lambda_1 \mathcal{L}_f + \lambda_2 \mathcal{L}_v + \lambda_3 \mathcal{L}_{veri} \tag{15}
\]

### 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 collected 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. In order to simulate to actual detect conditions, there are round 3,200 distractors sequences in the dataset.

For evaluation, we follow previous works and adopt the Cumulative Matching Characteristic (CMC) table and mean Average Precision (mAP) to evaluate the performance for all the datasets. For more details, we refer readers to the original paper [36]. In terms of iLIDS-VID and PRID2011, we only report the cumulative re-identification accuracy because that there only contain a single correct match in the gallery set.

4.2. Implementation Details

We implement our framework based on the Pytorch 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 256×128 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 mini-batch. 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-sample operation to increase the feature resolution. During training, we train our network for 50 epochs combining with the reciprocal learning approach. Considering that the low-correlated regions in a frame could be meaningful for identification, we design a novel temporal reciprocal learning mechanism to execute different learning strategies. “+ TRL” means that temporal reciprocal learning module with bi-directions is used to enhance and accumulate temporal information. In Tab. 1, we find that, compared with the “+ GCE”, the performance gets a slight drop when using F_{l} to test. This is due to the inhibition of most discriminative information in the low-correlation regions. However, the high-correlation feature F_{h} after reciprocal learning, has a better performance than + TRL.

Table 1. Ablation results of key components on MARS, iLIDS-VID and PRID2011.

| Methods  | Direction | MARS  | iLIDS-VID | PRID2011 |
|----------|-----------|-------|----------|----------|
|          |           | mAP Rank-1 | Rank-5 | Rank-20 |
| Baseline |           | 84.8   | 91.0     | 96.7     | 98.4     |
| + GCE    |           | 85.9   | 92.9     | 98.4     | 99.5     |
| + TRL    |           | 85.8   | 91.0     | 98.4     | 99.5     |

Table 2. Ablation results of the direction order.

| Methods | Direction | MARS  | iLIDS-VID | PRID2011 |
|---------|-----------|-------|----------|----------|
|         |           | Rank-1 | Rank-5   | Rank-20  |
| Baseline|           | 81.2   | 88.5     | 95.5     | 97.9     |
| GRL     | Forward   | 84.0   | 90.0     | 96.5     | 98.3     |
|         | Backward  | 83.7   | 90.0     | 96.5     | 98.3     |
|         | Bi-direction | 84.8 | 91.0     | 96.7     | 98.4     |

Table 3. Ablation results of the sequence length.

| Methods | Length | MARS  | iLIDS-VID | PRID2011 |
|---------|--------|-------|----------|----------|
|         |        | Rank-1 | Rank-5   | Rank-20  |
| Baseline| 8      | 81.2   | 88.5     | 95.5     | 97.9     |
| GRL     | 4      | 83.0   | 90.1     | 96.7     | 98.4     |
|         | 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     |

4.3. Ablation Study

In this subsection, we conduct experiments to investigate the effectiveness of proposed approach. The network is trained and evaluated on MARS, iLIDS-VID and PRID2011 datasets. The results are shown in Tab. 1, 2 and 3. In these tables, “Baseline” represents the backbone trained only with video-level OIM loss on the global branch. The direct spatial-temporal average pooling operation is applied on the multiple frame feature maps obtained from backbone. “GRL” represents our proposed global-guided reciprocal learning approach.

Effectiveness of Key Components. The ablation results of key components is reported Tab. 1 on three benchmark datasets. In this table, F_{g} denotes the global feature vector without correlation estimation is used to train with video-level OIM loss. F_{l} denotes the final feature vector responds to low-correlation maps also is supervised with video-level OIM loss. F_{h} denotes the final feature vector with respect to high-correlation maps is trained with frame-level OIM loss. Compared with “Baseline”, “+ GCE” means that we add the global-guided correlation estimation to guide the learning of spatial features. we can see that, when the the correlation maps are utilized, the performance have a significant improvement. The improvements indicate that it is beneficial to estimate the correlations in a global view.

Considering that the low-correlated regions in a frame could be meaningful for identification, we design a novel temporal reciprocal learning mechanism to execute different learning strategies. “+ TRL” means that temporal reciprocal learning module with bi-directions is used to enhance and accumulate temporal information. In Tab. 1, we find that, compared with the “+ GCE”, the performance gets a slight drop when using F_{l} to test. This is due to the inhibition of most discriminative information in the low-correlation regions. However, the high-correlation feature F_{h} after reciprocal learning, has a better performance than...
Figure 4. The visualization of the correlation maps and the channel activate maps of accumulative features at different steps. The top images are raw images from different time steps in video sequences. The heat maps at the second row are the high-correlation maps \( \{ R^t \}_{t=1}^{T=8} \) guided by the global view. The heat maps at the third and fourth rows are channel activate maps of the accumulative features, \( \{ M_1^t \}_{t=1}^{T=8} \) and \( \{ M_2^t \}_{t=1}^{T=8} \) in the forward and backward process.

only using the global-guided approach. It verified the effectiveness of semantic the designed enhancement block. Moreover, the combination of the low- and high-correlation feature vectors will boost the performance than either one does alone. Heavily, the increments benefit from jointly exploiting the conspicuous and informative characteristics and the sub-critical details. 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. The results demonstrate the effectiveness of our global-guided reciprocal learning.

**Influence of the Direction on Orders.** We perform experiments to investigate the effect of varying the direction of the temporal process. As shown in Tab. 2, the proposed temporal learning with forward or backward direction gains similar results. And the bi-directional reciprocating learning shows higher performances, which benefits from the combination between forward and backward temporal learning. The effectiveness shows that the aggregated features by reciprocating learning are more robust.

**Influence of the Different Sequence Lengths.** We train and test our bi-directional temporal reciprocating learning module with various sequence lengths \( T \), and the results are shown in Fig. 3. From this table, 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 will collect more meaningful cues with the increase of the sequence length. But excessive long sequences are bad for training the temporal reciprocal learning module.

**4.4. Visualization Analysis**

We visualize the learned high-correlation maps and the channel activate maps in Fig. 4. At the second row of this table, the high-correlation maps guided by the global view are shown. It can be obviously observed that, the high-correlation maps will focus on the foreground regions. The more significant in spatial and more consecutive in temporal the visual information is, the higher the correlation value is. For examples, the upper of human body usually has higher correlation values than the lower body. At the third and fourth rows, the channel activate maps of the accumulative features after a Res block, are visualized. 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 but meaningful cues, such as the shoes or bags, with red bounding boxes in Fig. 4. Meanwhile, we can find that, at the same time step, there are some variation among the features from the forward and backward process. which is useful to assemble more discriminative information for identification and improve the performance. The visual maps further validate that our method could highlight the most significant and aligned characteristics in temporal and pay more attentions on the sub-critical subjects on spatial, simultaneously.
Table 4. Comparison with state-of-the-art video-based person re-identification methods on MARS, iLIDS-VID and PRID2011.

| Methods    | Source | MARS          |                      |                      | iLIDS-VID          |                      |                      | PRID2011          |
|------------|--------|---------------|---------------------|---------------------|--------------------|---------------------|---------------------|-------------------|
|            |        | mAP | Rank-1 | Rank-5 | Rank-20 | mAP | Rank-1 | Rank-5 | Rank-20 | mAP | Rank-1 | Rank-5 | Rank-20 |
| SeeForest  | CVPR17 | 50.7 | 70.6 | 90.0  | 97.6   | 55.2 | 86.5 | 97.0   | 99.3   | 93.9 | 99.5 | 100   |         |
| ASTPN [30] | ICCV17 | -   | 44.0 | 70.0  | 81.0   | 62.0 | 86.0 | 98.0   | 99.0   | 77.0 | 95.0 | 99.0  |         |
| Snippet    | CVPR18 | 76.1 | 86.3 | 94.7  | 98.2   | 85.4 | 96.7 | 99.5   | 99.8   | 92.7 | 98.8 | 99.8  |         |
| STAN [16]  | CVPR18 | 65.8 | 82.3 | -     | -      | 80.2 | 93.2 | -      | -      | 94.4 | 100  | -     |         |
| STMP [19]  | AAAI19 | 72.7 | 84.4 | 93.2  | 96.3   | 84.3 | 96.8 | 99.5   | 99.8   | 92.7 | 98.8 | 99.8  |         |
| M3D [15]   | AAAI19 | 74.0 | 84.3 | 93.8  | 97.7   | 74.0 | 94.3 | -      | -      | 94.4 | 100  | -     |         |
| ST [7]     | AAAI19 | 80.8 | 86.3 | 95.7  | 98.1   | -    | -    | -      | -      | -    | -    | -     |         |
| Attribute  | CVPR19 | 78.2 | 87.0 | 95.4  | 98.7   | 86.3 | 97.4 | 99.7   | 100.0  | 93.9 | 99.5 | 100.0 |         |
| VRSTC [12] | CVPR19 | 82.3 | 88.5 | 96.5  | 97.4   | 83.4 | 95.5 | 99.5   | 100.0  | -    | -    | -     |         |
| GLTR [14]  | ICCV19 | 78.5 | 87.0 | 95.8  | 98.2   | 86.0 | 98.0 | -      | 100.0  | 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  | 98.5   | 88.6 | 98.0 | 99.7   | 100.0  | 95.9 | 99.7 | 100.0 |         |
| STGCN [31] | CVPR20 | 83.7 | 89.9 | -     | -      | -    | -    | -      | -      | -    | -    | -     |         |
| AFN [3]    | ECCV20 | 82.9 | 90.2 | 96.6  | -      | 88.5 | 96.8 | 99.7   | -      | -    | -    | -     |         |
| TCLNet [11] | ECCV20 | 85.1 | 89.8 | -     | -      | 86.6 | -    | -      | -      | -    | -    | -     |         |
| Ours       |        | 84.8 | 91.0 | 96.7  | 98.4   | 90.4 | 98.3 | 99.8   | 96.2   | 99.7 | 100  | -     |         |

4.5. Comparison with State-of-the-arts

In this section, the proposed approach is compared with state-of-the-art methods on three video-based person Re-ID benchmarks. The results are reported in Tab. 4. 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% on Rank-1 accuracy on iLIDS-VID dataset and PRID2011 dataset. We can see that the Rank-1 accuracy of our method outperforms all the compared methods, showing significant improvement over existing best state-of-the-art method. We notice that the MGRA [33] also employs the global view to aid the video-based person Re-ID task. It gains remarkable 85.9% mAP on MARS dataset. Different from MGRA using a reference representation for modeling the relations between global and all the node features, our method utilizes the global vectors to estimate two correlation maps for the feature disentanglement on spatial features. Jointly the proposed reciprocal learning, our proposed method could fully take the advantage of the disentangled features, and could explore more informative and fine-grained cues from the high- or low-correlation maps. Thereby, our method surpasses MGRA by 3.2%, 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 modeling the temporal information for video representations. STMP [19] introduces a refining recurrent unit to recover the missing parts by referring historical frames. In GLTR [14], they employ dilated temporal convolutions to capture the multi-granular temporal dependencies and aggregates obtained short and long-term temporal cues for global-local temporal representation. Particularly, compared with other methods for temporal learning, our proposed method achieves better results on three public datasets. Specifically, 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 to existing methods, our method utilizes the global information to guide the feature disentanglement. In addition, we adopt two strategies for mining 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.

5. Conclusion

In this paper, we propose a novel global-guided reciprocal learning framework for video-based person Re-ID. We present a GCE module to estimate the correlation degree of each pixel-level features in one frame. Then, the spatial features are disentangled into the high- and low-correlation features. To take full advantage of abundant visual information in sequences for identification, we propose a novel TRL module. In the temporal reciprocal learning module, multiple enhancement and memory unit is designed and connected in series for temporal learning. Besides, bidirectional process is arranged for forward and backward learning to resemble more robust representation. Based on the proposed modules, our approach could not only enhance the conspicuous and aligned information from the high-correlation features, but also accumulative sub-critical and fine-grained cues from the low-correlation features. Extensive experiments on public benchmarks show that our framework outperforms the state-of-the-arts.
References

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

[2] Daqin Feng, 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, 8

[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. Springer, 2020. 8

[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. 2

[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, pages --, 2019. 2, 6, 8

[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 Scandinavian Conference on Image Analysis, pages 91–102. Springer, 2011. 5

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):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. 8

[12] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinquan Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In CVPR, pages 7183–7192, 2019. 8

[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. 8

[15] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multiscale 3d convolution network for video-based person re-identification. In AAAI, volume 33, pages 8618–8625, 2019. 2, 8

[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. 1, 2, 3, 8

[17] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A spatio-temporal appearance representation for video-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, 8

[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, 8

[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. 8

[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, Shaoang Gong, Xiadian 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, 8

[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. 8

[32] Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *CVPR*, pages 6781–6789, 2018. 3

[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, 8

[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, 8

[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, 6

[36] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv:1610.02984*, 2016. 6

[37] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tie-niu 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. 8