The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Relation-Guided Spatial Attention and Temporal Refinement for Video-Based Person Re-Identification

Xingze Li, Wengang Zhou, Yun Zhou, Houqiang Li
CAS Key Laboratory of Technology in GIPAS, EEIS Department, University of Science and Technology of China
lixingze@mail.ustc.edu.cn, {zhwg, zhouyun, lihq}@ustc.edu.cn

Abstract

Video-based person re-identification has received considerable attention in recent years due to its significant application in video surveillance. Compared with image-based person re-identification, video-based person re-identification is characterized by a much richer context, which raises the significance of identifying informative regions and fusing the temporal information across frames. In this paper, we propose two relation-guided modules to learn reinforced feature representations for effective re-identification. First, a relation-guided spatial attention (RGSA) module is designed to explore the discriminative regions globally. The weight at each position is determined by its feature as well as the relation features from other positions, revealing the dependence between local and global contents. Based on the adaptively weighted frame-level feature, then, a relation-guided temporal refinement (RGTR) module is proposed to further refine the feature representations across frames. The learned relation information via the RGTR module enables the individual frames to complement each other in an aggregation manner, leading to robust video-level feature representations. Extensive experiments on four prevalent benchmarks verify the state-of-the-art performance of the proposed method.

1 Introduction

Person re-identification aims at matching the images of a person captured by multiple cameras, and in most cases the fields of view of these cameras are non-overlapping. It has a significant application in video surveillance and public security. This task is very challenging due to the variations of viewpoint, illumination and pedestrian's pose, as well as blur, occlusion and background clutter.

Depending on whether the data type is image or video, person re-identification is further divided into two subtasks, image-based person re-identification, and video-based person re-identification. In recent years, image-based person re-identification has achieved impressive progress (Sun et al. 2018; Fu et al. 2019b; Wang et al. 2018a; Hou et al. 2019a; Zhang et al. 2019; Zheng et al. 2019). However, this subtask is heavily influenced by the corrupted images where target blur or occlusion occurs. Compared with a single image with limited context, a video sequence captures abundant context information in a long span of time. In addition, it is more likely to find clean and informative content in a video, alleviating the noise sensitivity issue in image-based person re-identification. Therefore, how to explore this spatial and temporal information is the key to the video-based person re-identification.

To leverage such information, some state-of-the-art methods (Liu, Yan, and Ouyang 2017; Song et al. 2018; Fu et al. 2019a) estimate the qualities of global or local regions and use the qualities as weights to fuse features. Typically, these methods merely consider the per-region quality individually, which ignores the quality variance within a region and the context information. Intuitively, in the spatial domain, the operations within a local neighborhood fail to include enough positional information, while the global comparison is qualified to identify the valuable foreground and noisy background. Moreover, in the temporal domain, such a relation modeling can be naturally extended to multiple frames for frame-level complementation, which is beneficial for enhancing the frame-level representations. However, such global relation information is rarely explored by the existing video-based person re-identification methods.

In this paper, we propose two relation-guided modules to exploit the global relation information in both the spatial and the temporal domain. First, to capture the correspondence between two features, a relation module (RM) is designed to compute the relation vector. Then, to simultaneously localize the informative regions and depress the background globally, a relation-guided spatial attention (RGSA) module is developed. The attention at each position is determined by its feature and the relation vectors with all the positions, which is able to capture the local and global information. Finally, to further refine and enhance frame-level features, a relation-guided temporal refinement (RGTR) module is proposed. The relation information within all frames enables the individual frames to complement each other, contributing to reinforced frame-level representations, and so are the video-level representations. These two relation-guided modules cooperate in the spatial and temporal domain, leading to robust representations for person re-identification.

We conduct extensive experiments on four benchmarks,

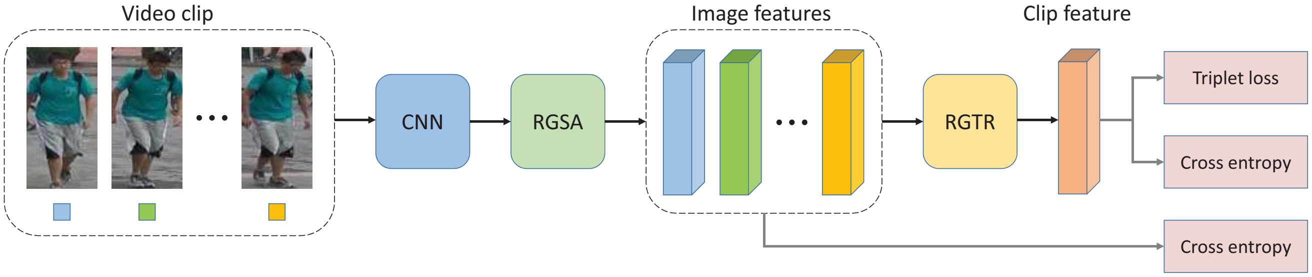

Figure 1: The overall architecture of our proposed method. The input video clip consists of T frames randomly sampled from a video sequence. The feature maps of these frames are extracted by a CNN backbone, and the relation-guided spatial attention (RGSA) module is deployed to generate frame-level features. All features of the clip are refined and fused using the relation-guided temporal refinement (RGTR) module. In the training phase, frame-level cross entropy loss and clip-level cross entropy loss are calculated on the corresponding features. Triplet loss is also computed on the clip-level features as metric learning.

and the experimental results demonstrate the state-of-the-art performance of our proposed method. Especially, to the best of our knowledge, our approach outperforms all state-of-the-art methods under multiple evaluation metrics on the large-scale MARS and DukeMTMC-VideoReID datasets.

2 Related Works

Video-based Person Re-Identification. Most works have focused on modeling the spatial and temporal information using recurrent neural networks, convolutional operations, and attention mechanisms.

Recurrent neural network is a common tool for sequence data processing. McLaughlin *et al.* introduce an RNN model and use a temporal average pooling operation to generate the video-level features in (McLaughlin, Martinez del Rincon, and Miller 2016). A multi-rate gated recurrent unit (GRU) is utilized in (Li et al. 2018b). Liu *et al.* use an RNN model to recover missing activation occurred in different regions and integrate both spatial and temporal clues in (Liu et al. 2019). With a similar goal, Hou *et al.* deploy a generative adversarial network to recover the occluded parts from adjacent frames in (Hou et al. 2019b).

Convolutional operation is another widely used method for video processing. In (Wu et al. 2018a), Wu *et al.* demonstrate that the temporal convolution network focuses more on the mid-level representation of motion, while optical flow captures the low-level motion information. A multi-scale 3D convolution network is used in (Li, Zhang, and Huang 2019) to learn the temporal cues.

Meanwhile, many works learn the attention mechanism to focus on more discriminative regions and frames. In (Xu et al. 2017), Xu *et al.* introduce a parameter matrix to capture attentive score in temporal dimension. Liu *et al.* estimate the quality score of each frame by a CNN model, and the video-level feature is the weighted sum of frame-level features in (Liu, Yan, and Ouyang 2017). A similar strategy is used in (Song et al. 2018; Fu et al. 2019a; Li et al. 2018a) to learn the qualities of multiple body regions in a frame. Chen *et al.* employ query and key-value projection to learn co-attentive embedding within two snippets in (Chen et al. 2018).

Different from these methods, where local information is captured, our approach utilizes the rarely explored global relation information to guide the spatial attention and temporal refinement.

Non-Local Mechanisms. Compared with local operations, the non-local mechanism explores global dependence, and it is used in many areas like video recognition, natural language processing, and object detection. In (Wang et al. 2018b), Wang *et al.* adopt non-local mean where the similarities between features are normalized to work as attention, and each feature is updated by the weighted sum of all features. In (Vaswani et al. 2017), the scaled dot product is performed between the query and all the keys, then, softmax is applied to obtain the weights on the values. Hu *et al.* consider both the appearance weight and the geometry weight to model the global relation between all objects in (Hu et al. 2018). The geometry weight makes sure that an object which satisfies specify geometry relation will have non-zero weight. In (Hou et al. 2019a), both non-local spatial aggregation and channel aggregation are deployed. Multi-context appearance relation is used to localize the body regions more precisely, and the location relation is inserted to constrain and complement the appearance relation.

In all the above non-local mechanisms, features are updated by a weighted sum operation, where similar features have large weights. Such methods suffer the limitations to capture context information effectively and identify the discriminative regions of the updated features. Different from these methods, our approach explores the global relation information to focus on the informative foreground and enable the frames to complement each other in the context.

Relation Reasoning. The Relation Network in (Santoro et al. 2017) considers the potential relations between all object pairs to capture the core common properties of relational reasoning. Temporal Relation Network accumulates multi-scale temporal relations to capture temporal relations in the video, where different temporal relation captures relationships between ordered frames of different length in (Zhou et al. 2018). Recurrent Relational Network performs multiple steps of relational reasoning and in each step, each object is affected by other objects, as well as its previous state in (Palm, Paquet, and Winther 2018).

3 Method

In this section, we first describe the overall architecture of our method. Then, the main components of our method are presented, including relation module, relation-guided spatial attention module, and relation-guided temporal refinement module. Finally, the loss function is discussed in detail.

3.1 Framework Overview

The framework of our proposed method is illustrated in Figure 1. Supposing the video clip is represented as $\{\mathbf{I}_t\}_{t=1}^{T}$, which contains T frames. For each frame, the CNN backbone is used to extract the feature maps $\mathbf{X}_t \in \mathcal{R}^{C\times H\times W}$, where C, H and W denote the channel, the height and the width of the feature maps respectively. Then, the feature maps are fed into the relation-guided spatial attention (RGSA) module to generate the frame-level feature $\mathbf{f}_t \in \mathcal{R}^{C}$. Finally, all features $\{\mathbf{f}_t\}_{t=1}^{T}$ of the video clip are refined through the relation-guided temporal refinement (RGTR) module to generate the clip-level feature $\tilde{\mathbf{f}} \in \mathcal{R}^{C}$. Frame-level cross entropy loss, clip-level cross entropy loss, and triplet loss are used together to optimize the model.

3.2 Relation Module

To calculate the relation between two features, the simplest way is to compute the inner product of these vectors. However, the inner product just indicates to what extent these features are similar. Some detailed information like which parts are similar and which parts are different can't be inferred from this relation. Another common relation is the element-wise difference. But difference relation is not compact and contains redundant information. It is also computation consuming for the following operations. Therefore, we develop a relation module to generate the relation vector of two features, which is both informative and compact compared with inner product and difference.

As illustrated in Figure 2, given two features $\mathbf{f}_1,\ \mathbf{f}_2$, we first compute the difference of the embedded features:

$$\mathbf{f}_{diff} = \theta(\mathbf{f}_1) - \phi(\mathbf{f}_2), \tag{1}$$

where θ and ϕ are two embedding functions implemented by a fully connected layer followed by a batch normalization (BN) and a rectified linear unit (ReLU), *i.e.,* $\theta(\mathbf{f}_1) = \mathrm{ReLU}(\mathrm{BN}(\mathbf{W}_\theta \mathbf{f}_1))$ and $\phi(\mathbf{f}_2) = \mathrm{ReLU}(\mathrm{BN}(\mathbf{W}_\phi \mathbf{f}_2))$. Fully connected layers $\mathbf{W}_\theta,\ \mathbf{W}_\phi \in \mathcal{R}^{\frac{C}{r_1}\times C}$ reduce the vector dimension with factor r_1. Then, we use a fully connected layer, a batch normalization layer and a rectified linear unit to generate more compact relation vector:

$$\mathbf{r}_{1,2} = \mathrm{RM}(\mathbf{f}_1,\ \mathbf{f}_2) = \mathrm{ReLU}(\mathrm{BN}(\mathbf{W}\mathbf{f}_{diff})), \tag{2}$$

where $\mathbf{W} \in \mathcal{R}^{\frac{C}{r_2}\times\frac{C}{r_1}}$, and the relation vector dimension is one r_2-th of the original vector.

3.3 Relation-Guided Spatial Attention

Deep stack of convolutional operations is often used to learn attention to focus on the foreground object. However, according to the study in (Luo et al. 2016), the effective receptive field is much smaller than the theoretical receptive field, *i.e.,* the attention is local-aware. To handle this problem, we develop a relation-guided spatial attention (RGSA) module. Attention at each spatial position is determined by its feature as well as the relation vectors from all positions, revealing the dependence between local and global information.

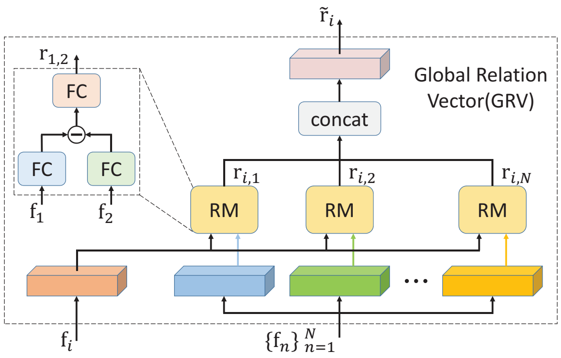

Figure 2: The architecture of the relation module (RM) and the global relation vector (GRV) module. The GRV module computes the relation vectors with N features and concatenates them into a global relation vector.

As illustrated in Figure 3(a), supposing a image's feature maps $\mathbf{X} \in \mathcal{R}^{C\times H\times W}$ are given, thus, there are N ($N = H \times W$) different spatial positions, and the feature at each position is a C-dimensional vector. Therefore, we first reshape the feature maps $\mathbf{X}$ to $\hat{\mathbf{X}} \in \mathcal{R}^{N\times C}$, and $\hat{\mathbf{X}}^i \in \mathcal{R}^{C} (1 \le i \le N)$ denotes the feature vector at i-th position. For each feature $\hat{\mathbf{X}}^i$, we use the relation module to compute its relation vectors with all features:

$$\mathbf{r}_{i,j} = \mathrm{RM}(\hat{\mathbf{X}}^i,\ \hat{\mathbf{X}}^j),\ (1 \le j \le N). \tag{3}$$

Then, all relation vectors related to $\hat{\mathbf{X}}^i$ are concatenated to generate the global relation vector:

$$\tilde{\mathbf{r}}_i = \mathrm{Concat}([\mathbf{r}_{i,1},\ \mathbf{r}_{i,2},\ \cdots,\ \mathbf{r}_{i,N}]), \tag{4}$$

where $\tilde{\mathbf{r}}_i \in \mathcal{R}^{\frac{NC}{r_2}}$ contains the global comparison information. Combined with the original feature, it can well guide the attention generation:

$$\mathbf{a}_i = \mathrm{Sigmoid}(\mathrm{BN}(\mathbf{W}_A[\hat{\mathbf{X}}^i,\ \tilde{\mathbf{r}}_i])), \tag{5}$$

where $\mathbf{W}_A \in \mathcal{R}^{C\times(\frac{N}{r_2}+1)C}$ and $\mathbf{a}_i \in \mathcal{R}^{C}$ has the same dimension with feature $\hat{\mathbf{X}}^i$. Finally, the frame-level feature is the weighted sum of the features of all positions as follows:

$$\mathbf{f} = \frac{\sum_{i=1}^{N} \mathbf{a}_i \hat{\mathbf{X}}^i}{\sum_{i=1}^{N} \mathbf{a}_i}. \tag{6}$$

3.4 Relation-Guided Temporal Refinement

Temporal feature fusion is the key point in video-based person re-identification. Some works (Liu, Yan, and Ouyang 2017) estimate the qualities of different frames and fuse the features by a weighted sum operation. But such a method still suffers the limitation that multiple low-quality frames

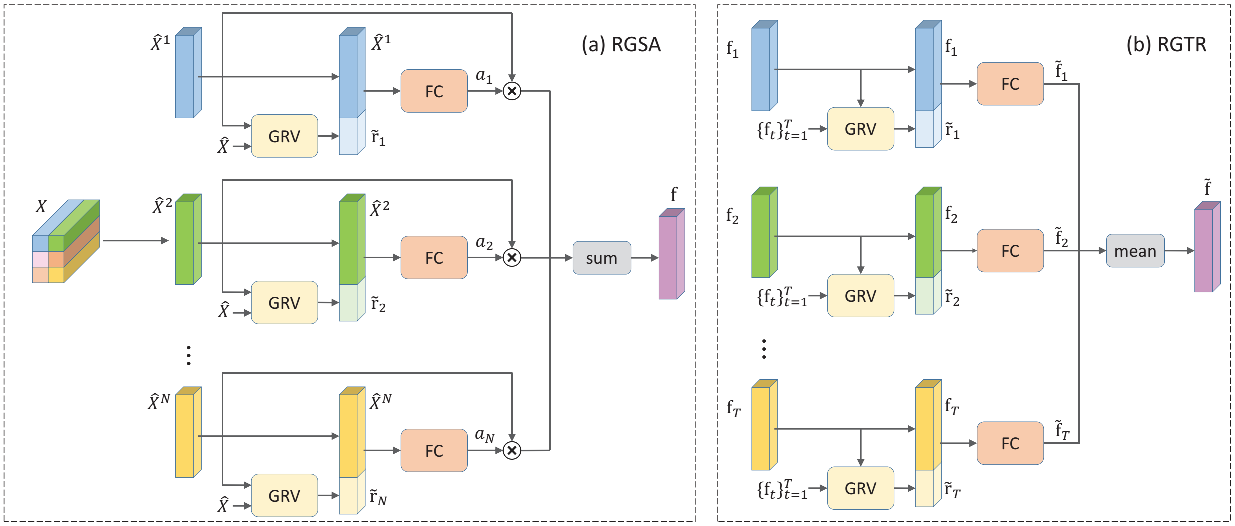

Figure 3: The architecture of the relation-guided spatial attention (RGSA) module and the relation-guided temporal refinement (RGTR) module. The GRV represents the global relation vector module as shown in Figure 2.

can be more informative when compared mutually. Different frames can complement each other and be refined and aggregated to enhance the discriminative capacity. Therefore, we develop a relation-guided temporal refinement (RGTR) module to refine the frame-level features by its relation with features of the other frames. Thus, each frame-level feature can be more robust and so is the clip-level feature.

As shown in Figure 3(b), by deploying the CNN backbone and the RGSA module, we can get frame features $\{\mathbf{f}_t\}_{t=1}^T$ of a video clip. Pairwise relation vectors between frames are computed using the relation module:

$$\mathbf{r}(t,s) = \text{RM}(\mathbf{f}_t,\ \mathbf{f}_s). \tag{7}$$

Then, all relation vectors related to $\mathbf{f}_t$ are concatenated to generate the global relation vector:

$$\tilde{\mathbf{r}}_t = \text{Concat}([\mathbf{r}(t,1),\ \mathbf{r}(t,2),\ \cdots,\ \mathbf{r}(t,T)]), \tag{8}$$

where $\tilde{\mathbf{r}}_t \in \mathcal{R}^{\frac{TC}{r_2}}$.

Different from learning attention in the RGSA module, which exploits the discriminative regions and depresses the background, the global relation vector is used in the RGTR module to enhance the discriminative capacity of all frame-level features, leading to reinforced clip-level representations. The refined frame-level features are dependent on the global relation vector and original features as formulated:

$$\tilde{\mathbf{f}}_t = \text{BN}(\mathbf{W}_R[\mathbf{f}_t,\ \tilde{\mathbf{r}}_t]), \tag{9}$$

where $\mathbf{W}_R \in \mathcal{R}^{C\times(\frac{T}{r_2}+1)C}$ and $\tilde{\mathbf{f}}_t \in \mathcal{R}^C$. Finally, the clip-level feature is obtained by temporal average pooling:

$$\tilde{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{f}}_t. \tag{10}$$

3.5 Loss Function

In our method, we adopt triplet loss, which is commonly used as metric learning, and cross entropy loss to train our model.

Each batch contains P identities and K video clips for each identity. And one video clip consists of T frames. Using the RGSA module and the RGTR module, we can extract frame-level features and clip-level features. Clip-level cross entropy loss $\mathcal{L}_{cce}$ is calculated following the setting of most methods. In addition, frame-level cross entropy loss $\mathcal{L}_{fce}$ is also employed to enhance the discriminative capacity of frame-level features, since, in most of the cases, the identity of the person can be determined by a single frame. Therefore, the total cross entropy loss $\mathcal{L}_{ce}$ is:

$$\mathcal{L}_{ce} = \mathcal{L}_{cce} + \mathcal{L}_{fce} \tag{11}$$

For each video clip $\tilde{\mathbf{f}}_{p,k}$ in the batch, we find its positive sets $\mathcal{P}_{p,k} = \{i|y_i = y_{p,k}\}$ and negative sets $\mathcal{N}_{p,k} = \{j|y_j \neq y_{p,k}\}$. Compared with batch hard triplet loss (Hermans, Beyer, and Leibe 2017) which chooses the hardest positive and negative samples, adaptive weighted triplet loss (Ristani and Tomasi 2018) uses all samples while hard samples have large weights and simple samples have small weights. The positive weight w_i and the negative weight w_i are computed as:

$$w_i = \frac{e^{D(\tilde{\mathbf{f}}_{p,k},\ \tilde{\mathbf{f}}_i)}}{\sum_{x\in\mathcal{P}_{p,k}} e^{D(\tilde{\mathbf{f}}_{p,k},\ \tilde{\mathbf{f}}_x)}}, \tag{12}$$

$$w_j = \frac{e^{-D(\tilde{\mathbf{f}}_{p,k},\ \tilde{\mathbf{f}}_j)}}{\sum_{x\in\mathcal{N}_{p,k}} e^{-D(\tilde{\mathbf{f}}_{p,k},\ \tilde{\mathbf{f}}_x)}}, \tag{13}$$

and the triplet loss $\mathcal{L}_{tri}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{tri} = \frac{1}{PK}\sum_{p,k}\text{softplus}(&\sum_{i\in\mathcal{P}_{p,k}}\omega_i D(\tilde{\mathbf{f}}_{p,k},\ \tilde{\mathbf{f}}_i) \\ &- \sum_{j\in\mathcal{N}_{p,k}}\omega_j D(\tilde{\mathbf{f}}_{p,k},\ \tilde{\mathbf{f}}_j)),\end{aligned} \tag{14}$$

where $\text{softplus}(x) = \ln(1+e^x)$ and $D(\cdot,\cdot)$ means the distance of two clip features.

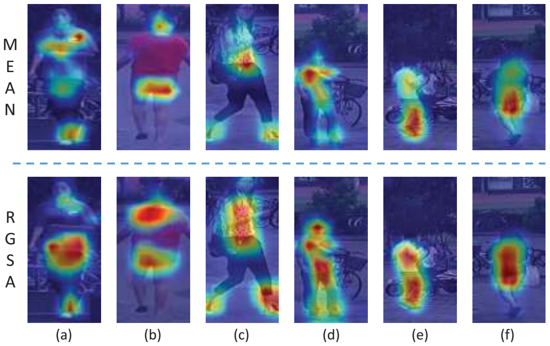

Figure 4: The Grad-CAM visualization results of example images from the MARS dataset. The images in the first row show the visualization results of the baseline model, and the images in the second row show the visualization results of the model with the RGSA module.

The overall training objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{tri}. \tag{15}$$

4 Experiments

In this section, we evaluate our method on four video-based person re-identification benchmarks. The datasets information and evaluation protocols, as well as experiment settings, are introduced first. Then, an ablation study is performed on the effectiveness of the components of our method. Finally, we compare our approach with the state-of-the-art methods.

4.1 Datasets and Evaluation Protocol

MARS (Zheng et al. 2016) is one of the largest video-based person re-identification benchmark with 1,261 identities and around 20,000 video sequences captured by six cameras in a university campus. The video sequences are detected and tracked by DPM detector (Felzenszwalb et al. 2008) and GMMCP tracker (Dehghan, Modiri Assari, and Shah 2015).

DukeMTMC-VideoReID (Wu et al. 2018b) is another large video-based person re-identification benchmark. It consists of 1,812 identities and 4,832 sequences captured by eight cameras. Each sequence has 168 frames on average. The bounding boxes are all manually annotated.

iLIDS-VID (Wang et al. 2014) and **PRID-2011** (Hirzer et al. 2011) are two small benchmarks. iLIDS-VID consists of 300 persons and each person has two sequences from two non-overlapping cameras, and the video sequences have an average length of 73. This dataset is challenging due to blur and occlusion. PRID-2011 contains 385 and 749 identities from two cameras respectively. Only the first 200 people appear in both cameras. Compared with iLIDS-VID, PRID-2011 has a relatively simple background and rare occlusion.

Evaluation Protocol. In our experiments, we use the Cumulative Matching Characteristic (CMC) and the mean average precision (mAP) to evaluate the performance. For the MARS and DukeMTMC-VideoReID datasets, we adopt the widely used training/testing splits provided by (Zheng et al. 2016) and (Wu et al. 2018b). For the iLIDS-VID and PRID-2011 datasets, we randomly split the identities equally into the training set and testing set. Experiments are conducted 10 times for average performance. Since each identity has just two sequences, only the CMC is used to evaluate the performance on the iLIDS-VID and PRID-2011 datasets.

Table 1: Ablation study on the components of the proposed method on the MARS and DukeMTMC-VideoReID datasets. The CMC score (%) at rank 1 and mAP (%) are reported.

Method	MARS		DukeMTMC -VideoReID	
	R1	mAP	R1	mAP
Baseline + $\mathcal{L}_{ce}$	87.2	80.9	95.3	94.2
Baseline + $\mathcal{L}_{total}$	87.8	81.8	95.4	94.5
RGSA + $\mathcal{L}_{total}$	88.6	83.2	96.0	95.0
RGTR + $\mathcal{L}_{total}$	89.4	83.3	96.0	95.3
RGSA+RGTR + $\mathcal{L}_{ce}$	88.5	83.5	96.3	95.3
RGSA+RGTR (IP) + $\mathcal{L}_{total}$	88.4	82.5	96.6	95.5
RGSA+RGTR + $\mathcal{L}_{total}$	89.4	84.0	97.2	95.8

4.2 Implementation Details

In the training phase, we randomly select T frames from a variable-length sequence to form a fixed-length input clip. Each batch consists of P identities and K input clips for each identity. In all our experiments, we select $P = 18$ and $K = 4$, therefore, the batch size is $72T$. All images are resized to 256×128, and randomly horizontal flipped. Random erasing (Zhong et al. 2017) is also used as data augmentation. We use the ResNet50 (He et al. 2016) pretrained on the ImageNet (Deng et al. 2009) dataset as backbone network. The last pooling layer and fully connected layer are removed and the stride in the last down-sampling in the conv5_x block is set to 1. The model is optimized using Adam (Kingma and Ba 2014) with weight decay 5×10^{-4}. The initial learning rate is 3×10^{-4} and it is reduced to 3×10^{-5} and 3×10^{-6} after training 125 and 250 epochs. The model is trained for 375 epochs in total. During the testing phase, the sequence is split into several video clips of length T with stride $T/2$. Clip-level features are extracted and L2-normalized. All clip-level features of the same sequence are averaged and L2-normalized to generate sequence-level features. Cosine distance is used to calculate the distance between query sequences and gallery sequences. Our model is implemented by Pytorch and optimized using four NVIDIA Tesla V100 GPUs.

4.3 Ablation Study

Effectiveness of Components. In Table 1, we investigate the effectiveness of the components of our method. In the **Baseline** model, the spatial and temporal average pooling operations are used to generate the frame-level and clip-level features. $\mathcal{L}_{ce}$ means that the training objective is both the frame-level entropy loss and clip-level cross entropy loss. $\mathcal{L}_{total}$ denotes that triplet loss and cross entropy loss

Table 2: Performance comparison with difference factors on the MARS and DukeMTMC-VideoReID datasets. The CMC scores (%) at rank 1 and mAP (%) are reported.

r_1	r_2	MARS		DukeMTMC -VideoReID	
		R1	mAP	R1	mAP
16	128	89.6	84.0	96.7	95.5
16	256	89.3	84.1	96.6	95.6
32	128	89.4	84.0	96.2	95.1
32	256	89.4	84.0	97.2	95.8

Table 3: Performance comparison with difference clip length on the MARS and DukeMTMC-VideoReID datasets. The CMC score (%) at rank 1 and mAP (%) are reported.

Clip Length	MARS		DukeMTMC -VideoReID	
	R1	mAP	R1	mAP
$T = 2$	88.5	82.5	95.9	94.9
$T = 4$	88.6	83.4	96.0	95.1
$T = 6$	89.0	83.5	96.2	95.1
$T = 8$	89.4	84.0	97.2	95.8

Table 4: Performance comparison with the state-of-the-art methods on the MARS dataset. The CMC scores (%) at rank 1, 5, 20 and mAP (%) are reported.

Method	MARS			
	R1	R5	R20	mAP
CNN+XQDA (Zheng et al. 2016)	68.3	82.6	89.4	49.3
SeeForest (Zhou et al. 2017)	70.6	90.0	97.6	50.7
RQEN (Song et al. 2018)	73.7	84.9	91.6	51.7
DuATM (Si et al. 2018)	78.7	90.9	95.8	62.3
ETAP (Wu et al. 2018b)	80.8	92.1	96.1	67.4
DRSTA (Li et al. 2018a)	82.3	-	-	65.8
Snippet (Chen et al. 2018)	86.3	94.7	98.2	76.1
RRU (Liu et al. 2019)	84.4	93.2	96.3	72.7
M3D (Li, Zhang, and Huang 2019)	84.4	93.8	97.7	74.1
STA (Fu et al. 2019a)	86.3	95.7	98.1	80.8
VRSTC (Hou et al. 2019b)	88.5	96.5	97.4	82.3
Ours	**89.4**	**96.9**	**98.3**	**84.0**

are used together to optimize the model. The **RGSA** and **RGTR** represent that the corresponding average pooling in the **Baseline** model is replaced with the proposed module. The **IP** represents that in the RGSA and RGTR modules, the relation module is replaced by the inner product to calculate the relation of two features.

Compared with **Baseline** + $\mathcal{L}_{total}$, the RGSA module improves rank-1 and mAP by 0.8% and 1.4% on MARS, as well as 0.6% and 0.5% on DukeMTMC-VideoReID. In Figure 4, we deploy the Grad-CAM (Selvaraju et al. 2017) to visualize the results for analyzing the effect of the RGSA module. The first row shows the results of the baseline model, and the second row shows the results of our RGSA module. We find that the RGSA module can help the model focus on the discriminative regions, and help cover the informative regions. In addition, the activations at these regions are stronger than the baseline method in Figure 4(a)(b)(c). In images that the bounding boxes are not well annotated like Figure 4(d)(e)(f), the pedestrian appears only in a small region of the image, the RGSA module also helps localize the pedestrian correctly which can be contributed to the exploiting of global comparison information in the module. These results show that our RGSA module is effective at localizing discriminative regions and depressing the background noise.

The employment of our RGTR module also boosts the performance compared with **Baseline** + $\mathcal{L}_{total}$, rank-1 and mAP improved by 1.6% and 1.5% on MARS, as well as 0.6% and 0.8% on DukeMTMC-VideoReID, since our RGTR module considers the influence of other frames. The comparison within all frames helps guide the modification and enhancement of each frame-level feature, thus improves the discriminative capacity of clip-level features. Compared with **RGSA** + $\mathcal{L}_{total}$ and **RGTR** + $\mathcal{L}_{total}$, the deployment of both the RGSA and RGTR modules can further improve the performance, rank-1 and mAP achieving 89.4% and 84.0% on MARS, as well as 97.2% and 95.8% on DukeMTMC-VideoReID. This result demonstrates that these modules can work cooperatively with each other in the spatial and temporal domain and contribute to superior feature representations.

Triplet loss is an effective tool for metric learning, which improves the performance of the **Baseline** model. Under the setting of without triplet loss, the use of the two relation-guided modules improves rank-1 and mAP by 1.3% and 2.6% on MARS, as well as 1.0% and 1.1% on DukeMTMC-VideoReID, compared with the **Baseline** + $\mathcal{L}_{ce}$ model. Such results also prove the effectiveness of these modules.

In **RGSA+RGTR (IP)** + $\mathcal{L}_{total}$, inner product is used to calculate the relation of two features, and such method outperforms the **Baseline** + $\mathcal{L}_{total}$ model. However, as discussed before, inner product can't provide detailed information and have limitation in guiding the spatial attention and temporal refinement. Compared with that, in **RGSA+RGTR** + $\mathcal{L}_{total}$, the deployment of our relation module improves the rank-1 and mAP by 1.0% and 1.5% on MARS, as well as 0.6% and 0.3% on DukeMTMC-VideoReID.

Influence of factors. In Table 2, we show the performance of the model with different settings of factors r_1 and r_2. All the experiments are trained with clip length of 8. We find that the rank-1 and mAP are very consistent with different settings, achieving about 89.5% and 84.0% on the MARS dataset, as well as 96.7% and 95.5% on the DukeMTMC-VideoReID dataset. These results show that relation module works effectively without the sensitivity of different choice of r_1 and r_2. In our final model, we adopt the setting with $r_1 = 32$ and $r_2 = 256$.

Influence of Clip Length. In Table 3, we investigate the influence of clip length. In the training phase, we randomly select T frames from a sequence to form an input video clip. As we can see, the performance improves grad-

Table 5: Performance comparison with the state-of-the-art methods on the iLIDS-VID and PRID-2011 datasets. The CMC scores (%) at rank 1, 5, 20 are reported.

Method	Ref	iLIDS-VID			PRID-2011		
		R1	R5	R20	R1	R5	R20
TDL (You et al. 2016)	CVPR	56.3	87.6	98.3	56.7	80.0	93.6
RNN (McLaughlin, Martinez del Rincon, and Miller 2016)	CVPR	58.0	84.0	96.0	70.0	90.0	97.0
CNN+XQDA (Zheng et al. 2016)	ECCV	53.0	81.4	95.1	77.3	93.5	99.3
SeeForest (Zhou et al. 2017)	CVPR	55.2	86.5	97.0	79.4	94.4	99.3
QAN (Liu, Yan, and Ouyang 2017)	CVPR	68.0	86.8	97.4	90.3	98.2	**100.0**
TSSCNN (Chung, Tahboub, and Delp 2017)	ICCV	60.0	86.0	97.0	78.0	94.0	99.0
ASTPN (Xu et al. 2017)	ICCV	62.0	86.0	98.0	77.0	95.0	99.0
SPW (Huang et al. 2018)	AAAI	69.3	89.6	98.2	83.5	96.3	**100.0**
RQEN (Song et al. 2018)	AAAI	76.1	92.9	99.3	92.4	98.8	**100.0**
TCN (Wu et al. 2018a)	AAAI	60.6	83.8	95.8	81.1	95.0	98.7
MGRU (Li et al. 2018b)	AAAI	60.8	89.2	**99.5**	78.4	94.8	99.4
DRSTA (Li et al. 2018a)	CVPR	80.2	-	-	93.2	-	-
Snippet (Chen et al. 2018)	CVPR	85.4	96.7	**99.5**	93.0	99.3	**100.0**
M3D (Li, Zhang, and Huang 2019)	AAAI	74.0	94.3	-	**94.4**	**100.0**	-
RRU (Liu et al. 2019)	AAAI	84.3	96.8	**99.5**	92.7	98.8	99.8
VRSTC (Hou et al. 2019b)	CVPR	83.4	95.5	**99.5**	-	-	-
Ours		**86.0**	**98.0**	99.4	93.7	99.0	**100.0**

Table 6: Performance comparison with the state-of-the-art methods on the DukeMTMC-VideoReID dataset. The CMC scores (%) at rank 1, 5, 20 and mAP (%) are reported.

Method	DukeMTMC-VideoReID			
	R1	R5	R20	mAP
ETAP (Wu et al. 2018b)	83.6	94.6	97.6	78.3
VRSTC (Hou et al. 2019b)	95.0	99.1	99.4	93.5
STA (Fu et al. 2019a)	96.2	99.3	99.6	94.9
Ours	**97.2**	**99.4**	**99.9**	**95.8**

ually as the clip length increases, rank-1 and mAP increasing from 88.5% and 82.5% to 89.4% and 84.0% on MARS, as well as from 95.9% and 94.9% to 97.2% and 95.8% on DukeMTMC-VideoReID. And the setting with a length of 8 achieves the best performance. Such improvement is attributed to the fact that the RGTR module uses the global comparison information to refine every features. When the clip gets longer, more comparison information will be available for the RGTR module to modify and enhance the features effectively. We believe that as the clip length increases, our model will achieve better results. However, more frames mean more computation and more GPU memory. Considering the limited computation resources, we adopt the setting with 8 frames in our final model, which also achieves competitive performance.

4.4 Comparison with State-of-the-Art Methods

The comparison of our proposed method with the state-of-the-art methods is shown in Table 4, Table 5, and Table 6. Our method outperforms the best existing methods on two large datasets, MARS and DukeMTMC-VideoReID in Table 4 and Table 6. Compared with VRSTC, our method achieves 0.9% and 1.7% improvement for rank-1 accuracy and mAP, respectively on the MARS dataset. On the DukeMTMC-VideoReID dataset, our method outperforms STA by 1.0% and 0.9% for rank-1 accuracy and mAP. Considering the high performance, these improvements are also appreciable. On two small datasets, iLIDS-VID and PRID-2011, our method also achieves comparable results. On the iLIDS-VID dataset, our method beats Snippet on rank-1 and rank-5 accuracy by 0.6% and 1.3%. Note that Snippet utilizes optical flow as an extra input, which is proved effective (McLaughlin, Martinez del Rincon, and Miller 2016; Xu et al. 2017) but not used in our method. On the PRID-2011 dataset, our approach performs slightly worse than M3D. As discussed above, PRID-2011 is too small and less challenging to explore the capacity of our model.

5 Conclusions

In this paper, we devote our efforts to video-based person re-identification and propose a novel framework with a relation-guided spatial attention (RGSA) module and a relation-guided temporal refinement (RGTR) module. In the RGSA module, the global comparison information helps localize the discriminative regions. In the RGTR module, the relation within frames is used to refine and enhance each frame's feature. These modules together contribute to superior feature representations for video-based person re-identification. Notably, our approach outperforms all existing state-of-the-art methods on large-scale MARS and DukeMTMC-VideoReID datasets.

Acknowledgements. This work was supported in part to Dr. Wengang Zhou by the National Key R&D Program of China under Grant 2018YFB1402600, NSFC under contract No. 61822208 and Youth Innovation Promotion Association CAS (No. 2018497), and in part to Dr. Houqiang Li by NSFC under contract No. 61836011.

References

Chen, D.; Li, H.; Xiao, T.; Yi, S.; and Wang, X. 2018. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*.

Chung, D.; Tahboub, K.; and Delp, E. J. 2017. A two stream siamese convolutional neural network for person re-identification. In *ICCV*.

Dehghan, A.; Modiri Assari, S.; and Shah, M. 2015. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Felzenszwalb, P. F.; McAllester, D. A.; Ramanan, D.; et al. 2008. A discriminatively trained, multiscale, deformable part model. In *CVPR*.

Fu, Y.; Wang, X.; Wei, Y.; and Huang, T. 2019a. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*.

Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; and Huang, T. 2019b. Horizontal pyramid matching for person re-identification. In *AAAI*.

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

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. In *SCIA*.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019a. Interaction-and-aggregation network for person re-identification. In *CVPR*.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019b. Vrstc: Occlusion-free video person re-identification. In *CVPR*.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *CVPR*.

Huang, W.; Liang, C.; Yu, Y.; Wang, Z.; Ruan, W.; and Hu, R. 2018. Video-based person re-identification via self paced weighting. In *AAAI*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S.; Bak, S.; Carr, P.; and Wang, X. 2018a. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*.

Li, Z.; Yao, L.; Nie, F.; Zhang, D.; and Xu, M. 2018b. Multi-rate gated recurrent convolutional networks for video-based pedestrian re-identification. In *AAAI*.

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*.

Liu, Y.; Yuan, Z.; Zhou, W.; and Li, H. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*.

Liu, Y.; Yan, J.; and Ouyang, W. 2017. Quality aware network for set to set recognition. In *CVPR*.

Luo, W.; Li, Y.; Urtasun, R.; and Zemel, R. 2016. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*.

McLaughlin, N.; Martinez del Rincon, J.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *CVPR*.

Palm, R.; Paquet, U.; and Winther, O. 2018. Recurrent relational networks. In *NeurIPS*.

Ristani, E., and Tomasi, C. 2018. Features for multi-target multi-camera tracking and re-identification. In *CVPR*.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NeurIPS*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Si, J.; Zhang, H.; Li, C.-G.; Kuen, J.; Kong, X.; Kot, A. C.; and Wang, G. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*.

Song, G.; Leng, B.; Liu, Y.; Hetang, C.; and Cai, S. 2018. Region-based quality estimation network for large-scale person re-identification. In *AAAI*.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018a. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *CVPR*.

Wu, Y.; Qiu, J.; Takamatsu, J.; and Ogasawara, T. 2018a. Temporal-enhanced convolutional network for person re-identification. In *AAAI*.

Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; and Yang, Y. 2018b. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*.

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

You, J.; Wu, A.; Li, X.; and Zheng, W.-S. 2016. Top-push video-based person re-identification. In *CVPR*.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2019. Densely semantically aligned person re-identification. In *CVPR*.

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

Zheng, F.; Deng, C.; Sun, X.; Jiang, X.; Guo, X.; Yu, Z.; Huang, F.; and Ji, R. 2019. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2017. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*.

Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; and Tan, T. 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*.

Zhou, B.; Andonian, A.; Oliva, A.; and Torralba, A. 2018. Temporal relational reasoning in videos. In *ECCV*.