Frame-Guided Region-Aligned Representation for Video Person Re-Identification

Zengqun Chen, Zhiheng Zhou,* Junchu Huang, Pengyu Zhang, Bo Li
South China University of Technology, Guangzhou, China
{eechenzq, eehjc, eezhangpy}@mail.scut.edu.cn, {zhouzh, leebo}@scut.edu.cn

Abstract

Pedestrians in videos are usually in a moving state, resulting in serious spatial misalignment like scale variations and pose changes, which makes the video-based person re-identification problem more challenging. To address the above issue, in this paper, we propose a Frame-Guided Region-Aligned model (FGRA) for discriminative representation learning in two steps in an end-to-end manner. Firstly, based on a frame-guided feature learning strategy and a non-parametric alignment module, a novel alignment mechanism is proposed to extract well-aligned region features. Secondly, in order to form a sequence representation, an effective feature aggregation strategy that utilizes temporal alignment score and spatial attention is adopted to fuse region features in the temporal and spatial dimensions, respectively. Experiments are conducted on benchmark datasets to demonstrate the effectiveness of the proposed method to solve the misalignment problem and the superiority of the proposed method to the existing video-based person re-identification methods.

Introduction

The person re-identification research has drawn increasing attention in the computer vision field in recent years, as this problem underpins various critical applications such as surveillance, activity analysis and tracking. Given a target person appearing in a surveillance camera, this task aims to identify a set of matching person images from a pool, which are usually captured from non-overlapping cameras. It remains a challenging task due to the influence of cluttered background, occlusion, heavy illumination changes, non-rigid deformation of human bodies and viewpoint variations across camera views.

In recent years, much attention has been shifted to video-based re-ID because of its natural settings and impressive benefits of sequential information. In this paper, we study the person re-ID problem in the video setting. Many existing video-based person re-identification methods extract frame-level features and generate sequence-level features through average or maximum pooling (Liu et al. 2017; You et al. 2016; McLaughlin, Rincon, and Miller 2016; Zhou et al. 2017). In the presence of partial occlusion, the extracted sequence-level features are usually corrupted due to the equal treatment of all the frames, leading to severe performance degradation. To address this, recent works (Liu, Yan, and Ouyang 2017; Song et al. 2018; Li et al. 2018b) have proposed several impressive methods that generate the adaptive scores for feature weighting. Region-based methods (Song et al. 2018; Li et al. 2018b) concentrate more attention on the discriminative image regions and aggregate the complementary region information across frames. However, aggregating region features is not straightforward. As illustrated in Fig. 1, due to two types of detection errors: excessive background and body misalignment, poor spatial alignment of moving pedestrians deteriorates the quality of the extracted sequence-level features and compromise pedestrian matching. To this end, we propose a novel video representation learning scheme called Frame-Guided Region-Aligned model (FGRA) to effectively solve the misalignment problem.

Figure 1: Sample video sequences with misalignment. Body parts correspond to the excessive background and other body parts in the same spatial regions due to detection errors.

Our approach is partially motivated by the success of a frame-guided feature learning strategy in the video object segmentation field (Caelles et al. 2017; Perai et al. 2017) as well as the successful application of cross correlation to target location in the visual object tracking field (Bertinetto et al. 2016; Li et al. 2018a; 2019). One apparent distinction is that instead of using previous frame masks, we use discriminative feature vectors to pass guidance information across frames. Another distinction is that without predefined target objects, our approach automatically learns target tem-

*Corresponding author

plate features for region alignment. Specifically, we design our model as a two-branch architecture including a global branch and a local branch. The global branch is deployed to extract global spatiotemporal features for complementary discriminability. The local branch extracts local spatiotemporal features in two steps, including a region alignment mechanism and a feature aggregation strategy. The whole model is trained in an end-to-end manner.With the assist of the proposed region alignment mechanism and feature aggregation strategy, FGRA can drive the alignment of the corresponding regions across frames at the feature level, and explore more discriminative spatiotemporal cues that are robust to misalignment like scale variances and pose changes.

The main contributions in this paper can be summarized in four folds as follows:

- To the best of our knowledge, this is the first attempt to introduce the frame-guided alignment strategy for video-based person re-identification. We utilize well-learned region features of a reference frame as template features to explicitly align body regions across frames, which aims to explore more accurate spatiotemporal cues.
- We propose a novel non-parametric alignment module based on depth-wise cross correlation to generate well-aligned region features. A consistency regularization term is optimized to further ensure the compactness of the aligned features.
- We design temporal alignment score for temporal feature aggregation based on how relevant the region-aligned feature is to the template feature. We also adopt spatial attention to explore discriminative region information for spatial feature aggregation.
- We conduct extensive experiments to demonstrate the effectiveness of each component. The final results achieve significant performance compared to the existing state-of-the-art approaches on three mainstream video-based re-ID datasets: iLIDS-VID, PRID and MARS.

Related Work

In this section, we review the related works related to the video-based person re-identification, the frame-guided feature learning strategy and the cross correlation module.

Video-based Person Re-identification. Compared with image-based re-ID, video-based re-ID provides richer visual information and is promising for more accurate retrieval (Zheng et al. 2016). McLaughlin et al. (McLaughlin, Rincon, and Miller 2016) and Wu et al. (Wu, Shen, and Hengel 2016) built a CNN to extract per-frame feature and then utilized an RNN and temporal pooling for feature aggregation. RNN-based methods treat all the frames equally so that the poor-quality frames will distort the video representation. In order to better distill relevant information from the video, attention-based approaches are gaining popularity (Zhou et al. 2017; Xu et al. 2017; Liu, Yan, and Ouyang 2017; Li et al. 2018b; Song et al. 2018). Some recent works (Li et al. 2018b; Song et al. 2018) tend to learn region-based video representations and further tackle the misalignment problem. For example, Li et al. (Li et al. 2018b) utilized multiple spatial attention modules to roughly localize distinctive body parts of a person for region alignment. However, spatial attention may not steadily focus on the same regions and generate false saliency due to frequent detection errors. To alleviate this dilemma, we propose a novel region alignment mechanism based on a frame-guided feature learning strategy and a non-parametric alignment module to effectively solve the misalignment problem. Recently Hou et al. (Hou et al. 2019) proposed VRSTC to recover the occluded regions to train re-ID network, which can be treated as data preprocessing. The main contributions of our paper and VRSTC work on two different aspects and can be combined. In this paper, we do not concentrate on it since our work focuses on discriminative representation learning.

Frame-Guided Feature Learning Strategy. Consecutive frames contain strong information correlation between each other, and previous frame provides guidance for information analysis of subsequent frames, which has been practically accepted in the video object segmentation field (Caelles et al. 2017; Perai et al. 2017). Caelles et al. (Caelles et al. 2017) proposed OSVOS that utilized merely one mark of the first frame but gave rise to highly accurate and temporally consistent segmentations. Our work is partially motivated by this strategy. Corresponding body regions of a person in a sequence are highly similar to each other. It is desirable to locate and align the same body regions of the remaining frames based on the previous frame.

Cross Correlation Module. The cross correlation module is a significant operation to aggregate information of two objects, which has been successfully applied in visual tracking tasks (Bertinetto et al. 2016; Li et al. 2018a; 2019). For example, a unified framework SiamRPN++ (Li et al. 2019) presented a lightweight cross correlation layer called depth-wise cross correlation, which achieved efficient information association. Inspired by the above methods, our method extends them to extract region-aligned features.

Proposed Method

In this section, we introduce the overall system pipeline of the proposed method and then explain specific configuration of its important components in more detail.

Problem Formulation

Person re-identification can be implemented as a ranking task. Given a probe video sequence $\mathcal{Q} = \{q_n | q_n \in \mathbb{R}^D\}_{n=1}^N$, where N is the sequence length and D is the dimension of an image, the video-based re-ID task aims to retrieve the same person by ranking gallery sequences based on the similarity between the probe sequence $\mathcal{Q}$ and each gallery sequence $\mathcal{G} = \{g_n | g_n \in \mathbb{R}^D\}_{n=1}^N$. In the final search result list, the video sequences of the same person as the probe $\mathcal{Q}$ are assigned the top rank, i.e. the highest ranking score.

Architecture Overview

This work proposes a new deep learning architecture to learn region-aligned video representations guided by well-learned template features in an end-to-end manner. As illustrated in Fig. 2, an input video $\mathcal{S} = \{x_1, x_2, \cdots, x_N\}$ is fed into the backbone network to extract the frame features $\{\mathbf{F}_n | \mathbf{F}_n \in$

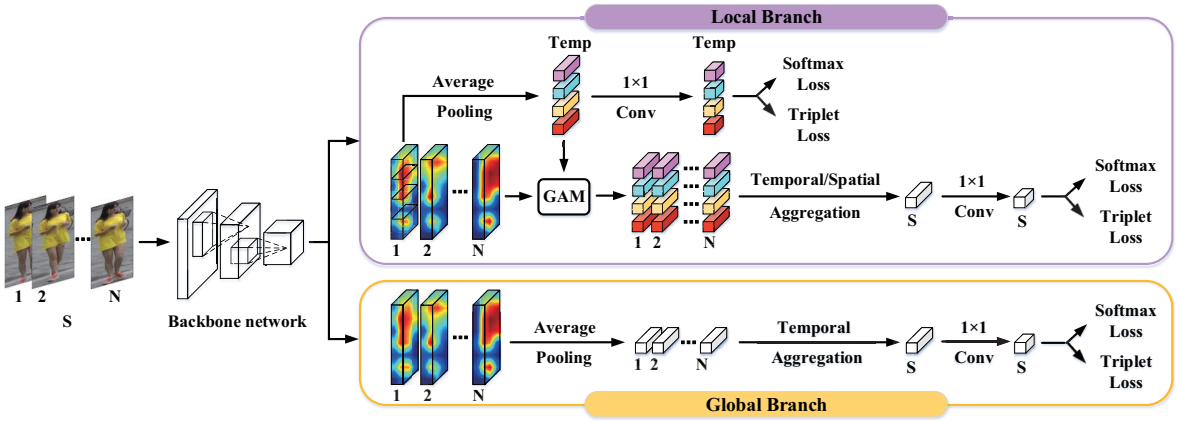


Figure 2: The architecture of FGRA framework. This framework is designed as a two-branch architecture. The global branch is deployed to extract global spatiotemporal features, and the local branch extracts local spatiotemporal features in two steps, including region alignment and feature aggregation. Specifically, the local branch is composed of multiple functional components including a template feature learning branch, a guide alignment module (GAM) and a spatial-temporal aggregation operation.

$\mathbb{R}^{C\times H\times W}\}_{n=1}^{N}$, where C, H and W denote the number of channels, height and width of the feature maps, respectively. We select Resnet50 (He et al. 2016) as the backbone network, where the global average pooling and fully connected layers are discarded. Then the frame features are processed through the global branch and local branch to obtain the global and local sequence features $[\mathbf{f}_{s,g} \in \mathbb{R}^c, \mathbf{f}_{s,l} \in \mathbb{R}^c]$, respectively, where c denotes the length of the reduced dimension. Both of them are complementary to each other and concatenated as the final representation to perfect the comprehensiveness for retrieval.

In the global branch, we first apply a global average pooling layer to each frame feature map, which is followed by a temporal attention module originally proposed in (Li et al. 2018b) to temporally aggregate the global features and generate a compact video representation. A 1×1 convolutional layer is further applied to reduce the feature dimension from 2048 to1024 and output the final representation $\mathbf{f}_{s,g}$.

In the local branch, we adopt a region alignment mechanism and a feature aggregation strategy to drive region alignment and capture discriminative spatiotemporal cues. We deploy a guide branch to learn template region features $\{\mathbf{F}_{t,i}|\mathbf{F}_{t,i} \in \mathbb{R}^C\}_{i=1}^{N_s}$ of a reference frame selected from a sequence of images, where N_s indicates the number of regions. Then the template features $\{\mathbf{F}_{t,i}\}_{i=1}^{N_s}$ and the frame features $\{\mathbf{F}_n\}_{n=1}^{N}$ are fed into a guide alignment module (GAM) to generate the region-aligned features $\{\{\mathbf{F}_{n,i}|\mathbf{F}_{n,i} \in \mathbb{R}^C\}_{i=1}^{N_s}\}_{n=1}^{N}$. The final local representation $\mathbf{f}_{s,l}$ is obtained by spatiotemporal weighted averaging and dimension reduction.

Region Alignment Mechanism

Pedestrians in most datasets are well-aligned by hand-drawn bounding boxes. But in reality, the bounding boxes of pedestrians are detected rather than manually labeled, and thus pedestrian matching may succumb to heavy misalignment and strong deformation. To this end, we propose an effective approach called region alignment mechanism, which automatically solves two common problems in video re-ID: aligning corresponding body regions across frames and determining which region is more informative.

Template Feature Learning. We treat the feature alignment challenge as a guided feature learning problem, considering that the previous frame can provide information clues for feature learning of subsequent consecutive frames. In the guide branch, we take the first frame of the input video sequence as a reference frame and spatially downsample the corresponding backbone feature into N_s column vectors $\{\mathbf{F}_{t,i}\}_{i=1}^{N_s}$. Afterward, a 1×1 kernel-sized convolution layer is employed to reduce the dimension of $\mathbf{F}_{t,i}$, which is independently optimized by the objective function.

Guide Alignment Module (GAM). In order to solve the misalignment problem, a novel non-parametric alignment module is embedded to the proposed architecture to achieve region alignment. As shown in Fig. 3, the template region feature $\mathbf{F}_{t,i}$ and the frame feature $\mathbf{F}_n$ with the same number of channels do the depth-wise cross correlation to produce the similarity maps $\mathbf{M}_{n,i} \in \mathbb{R}^{C\times H\times W}$, where the target region will get high response value. The above process can be formulated as a parameter-free function $f(\mathbf{F}_{t,i}, \mathbf{F}_n)$:

$$\mathbf{M}_{n,i} = f(\mathbf{F}_{t,i}, \mathbf{F}_n) = \mathbf{F}_{t,i} * \mathbf{F}_n, \tag{1}$$

where $*$ indicates group convolution operation. It is clear that depth-wise cross correlation is mathematically equivalent to group convolution and the template feature $\mathbf{F}_{t,i}$ can be regarded as the convolution kernel exactly.

We apply a batch normalization layer and a sigmoid function to the similarity maps to normalize each element to the range of (0,1). Then the frame feature $\mathbf{F}_n$ and the similarity maps $\mathbf{M}_{n,i}$ perform Hadamard product to enhance the feature saliency of the target region. The result of the calculation is further downsampled to obtain the final region-aligned feature $\mathbf{F}_{n,i}$, which can be summarized as :

$$\mathbf{F}_{n,i} = \frac{1}{H \times W} \sum^{H} \sum^{W} \mathbf{F}_n \circ \mathbf{M}_{n,i} \tag{2}$$

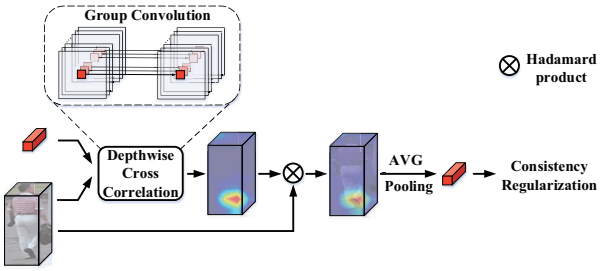


Figure 3: The details of the guide alignment module (GAM).

where $\circ$ denotes Hadamard product.

Consistency Regularization. The outlined method for region alignment might be subject to extreme light condition and superposition, which result in the false high saliency region and inaccurate spatiotemporal cues. In practice, we need to ensure the consistency of the aligned regions and avoid the case in which similar but misaligned regions gain high response value in the similarity maps.

Specifically, we design a regularization term based on center loss (Wen et al. 2016) to encourage the proposed model to maintain the alignment consistency. Typically, the regularization term for each video sequence is defined as:

$$L_{center} = \sum_{i=1}^{N_s} \sum_{n=1}^{N} \|\mathbf{F}_{n,i} - \mathbf{c}_i\|_2^2, \tag{3}$$

where $\mathbf{c}_i$ indicates the center of the region-aligned features $\{\mathbf{F}_{n,i}\}_{n=1}^{N}$. This regularization term is minimized to increase the compactness of the region-aligned features.

Feature Aggregation Strategy

As we obtain the region-aligned features, a video sequence $\mathcal{S}$ can be mapped to a local video representation $\mathbf{f}_{s,l}$ by spatial and temporal feature aggregation.

Temporal Alignment Score. To achieve temporal aggregation, we propose temporal alignment scores to conduct the weighted combination on the aligned features. Based on the similarity between the aligned feature and the template feature, the temporal alignment score is generated by feeding the similarity maps into a global average pooling layer and a softmax layer:

$$\hat{\mathbf{s}}_{n,i} = \frac{exp(\mathbf{s}_{n,i})}{\sum_{k=1}^{N} exp(\mathbf{s}_{k,i})}, \tag{4}$$

$$\mathbf{s}_{n,i} = \frac{1}{H \times W} \sum^{H} \sum^{W} \mathbf{M}_{n,i}. \tag{5}$$

Note that $\mathbf{s}_{n,i} \in \mathbb{R}^C, \hat{\mathbf{s}}_{n,i} \in \mathbb{R}^C$ are vectors instead of scalers, and Eq. 4 serves as a normalization function like a traditional softmax function but it is implemented along the time dimension. The temporal alignment score indicates how relevant the aligned feature is to the well-learned template feature. In other words, it reflects how informative the aligned features are to represent a pedestrian.

The region video representation $\mathbf{F}_{s,i} \in \mathbb{R}^C$ is obtained by fusing the region-aligned features in a weighted average manner:

$$\mathbf{F}_{s,i} = \sum_{n=1}^{N} \mathbf{F}_{n,i} \circ \hat{\mathbf{s}}_{n,i}, \tag{6}$$

where $\circ$ means Hamamard product.

Spatial Attention Feature Learning. Not all regions of an object are discriminative in every video frame because of similar clothes or explicit foreground occluders. Therefore, the effective region information is easily weakened by noise region information due to the equal treatment of all regions. To address this, we conduct the weighted combination over the region video representations for spatial aggregation. Specifically, we concatenate the region video representations $\{\mathbf{F}_{s,i}\}_{i=1}^{N_s}$ as $\mathbf{F}_s \in \mathbb{R}^{C \times N_s}$ and send it into a convolutional layer to generate spatial attention weights $\mathbf{W} \in \mathbb{R}^{1 \times N_s}$. The above process can be formulated as:

$$\mathbf{W} = softmax(g(\mathbf{F}_s)), \tag{7}$$

where $g(\cdot)$ denotes a series of operations including a 1D convolution layer with a filter of kernel size 3, a batch normalization layer and a ReLU unit. The softmax function is also used to normalize the spatial weights. Then the local video representation $\mathbf{F}_{s,l} \in \mathbb{R}^C$ can be calculated as follow:

$$\mathbf{F}_{s,l} = \mathbf{F}_s \cdot \mathbf{W}^{\mathrm{T}} \tag{8}$$

After dimension reduction on the local representation $\mathbf{F}_{s,l}$, the input video is represented by a single feature vector $\mathbf{f}_s$ generated by concatenating the global and local representations:

$$\mathbf{f}_s = [\mathbf{f}_{s,g}, \mathbf{f}_{s,l}] \tag{9}$$

Objective Functions

In this paper, we train our model in an end-to-end manner by optimizing the objective function composed of an identification loss function and a regularization term.

The identification loss L_{id} consists of softmax loss $L_{softmax}$ and triplet loss $L_{triplet}$, aiming to supervise the learning of discriminative representations of pedestrians. Typically, the original triplet loss proposed in (Hermans, Beyer, and Leibe 2017) is given by:

$$L_{triplet} = \sum_{i=1}^{P} \sum_{a=1}^{K} [m + \overbrace{\max_{p=1\ldots K} ||\mathbf{f}_a^i - \mathbf{f}_p^i||_2}^{hardest\ positive} - \underbrace{\min_{\substack{n=1\ldots K \\ j=1\ldots P \\ j \neq i}} ||\mathbf{f}_a^i - \mathbf{f}_n^j||_2}_{hardest\ negative}]_+, \tag{10}$$

where $\mathbf{f}_a^i$, $\mathbf{f}_p^i$ and $\mathbf{f}_n^j$ indicate the features extracted from anchor, positive and negative samples respectively, and m denotes the margin value between intra-class distance and inter-class distance. The original softmax loss can be formulated as follows:

$$L_{softmax} = -\frac{1}{P \times K} \sum_{i=1}^{P} \sum_{a=1}^{K} \log p(y_{i,a}|x_{i,a}) \tag{11}$$

where $y_{i,a}$ is the ground truth identity of the corresponding input tracklet $x_{i,a}$, and $p(y_{i,a}|x_{i,a})$ indicates the prediction probability of the classifier for the ground truth identity.

As mentioned, we design a regularization term to regularize the consistency of the aligned regions across frames, which can be globally given by:

$$L_{reg} = \frac{1}{P \times K} \sum^{P} \sum^{K} L_{center} \tag{12}$$

We multiply it by the coefficient λ and add it to the final objective function, which is minimized to optimize our model:

$$min(L_{id} + \lambda L_{reg}) \tag{13}$$

Experiments

Datasets and Protocols

We evaluate our approach on three well-known datasets, including **iLIDS-VID** (Wang et al. 2014), **PRID-2011** (Hirzer et al. 2011) and **MARS** (Zheng et al. 2016). 1) iLIDS-VID is very challenging due to clothing similarities among people and random occlusion. It consists of 600 image sequences of 300 people in total, recorded by a pair of non-overlapping cameras. Video sequences have various lengths ranging from 23 to 192 frames with an average duration of 73 frames. 2) PRID-2011 is captured in relatively simple environments with rare occlusion. It contains 749 persons captured by two cameras but only 200 persons appear in both cameras, constituting of 400 video sequences. The length of each video sequence varies from 5 to 675. Following Zheng et al. (Zheng et al. 2016), sequences with more 21 frames are selected, leading to 178 identities. 3) MARS is one of the largest video-based person re-identification datasets which contains 17503 tracklets from 1261 identities and additional 3248 tracklets of poor quality serving as distracters. Each identity is captured by at least 2 cameras and has 13.2 tracklets on average.

Evaluation Protocols. We use the standard experimental protocols for testing. For iLIDS-VID and PRID-2011 datasets, we follow the implementation in previous works (Wang et al. 2014) and randomly split the datasets into 50% of persons for training and 50 % of persons for testing. This procedure is repeated 10 times to calculate the average accuracy. The MARS dataset provides fixed training and testing sets, which contain predefined 8298 sequences of 625 persons for training and 12180 sequences of 636 persons for testing, including 3248 low-quality sequences in the gallery set. We employ CMC curve (Bolle et al. 2005) and mAP (Zheng et al. 2015) to evaluate the performance in our experiments. For ease of comparison, we only report the cumulated re-identification accuracy at the selected ranks.

Implementation Details

Since the video tracklet has a variable length, a subsequence of N = 6 frames is randomly selected from the entire tracklet as an input during the training process. The input size of each frame is set as 256×128 pixels, randomly cropped from a scaled image whose size is enlarged by $1/8$. We then apply the image-level augmentation to the whole sequence, including mirroring, normalization and randomly erasing (Zhong et al. 2017b). In order to train hard mining triplet loss, 16 identities with 4 tracklets each person are taken in a mini-batch so that the mini-batch size is 64. The number of spatial regions N_s in the local branch is set to 4. For better optimization of our model, we recommend to set the margin parameter in triplet loss to 0.5 and set the coefficient associated with center loss to 5e-4.

For network parameter training, we adopt Adam with a weight decay of 0.0005. The model is trained for 300 epochs in total, starting with a learning rate of 0.03 for parameters in the center loss and 0.0003 for others. The learning rate is reduced ten times after every 100 epochs.

In the test phase, in order to make full use of the information of the whole sequence, we segment several video clips from each sequence by the stride of 3, and every two adjacent clips overlap by half the length. The video clips pass forward the model to obtain the retrieval features. These features then conduct average pooling to compute the similarity based on the cosine distance. Note that different clips have different reference frames, which can effectively relieve the negative impact if the reference frame is of poor quality.

Comparison with State-of-the-art Approaches

We compare our proposed method with the state-of-the-art methods on three datasets: iLIDS-VID, PRID-2011 and MARS in Table 1. All the result are achieved without any post-processing techniques such as re-ranking (Zhong et al. 2017a) or multi-query (Zheng et al. 2016). As shown in Table 1, our method achieves the top-1 accuracy of **88.0%**, **95.5%** and **87.3%** on iLIDS-VID, PRID2011 and MARS, and the mAP of **81.2%** on MARS, outperforming the best performance of the published methods including extracting handcrafted features and learning deep features. The main reason for the improvements is that we propose a novel region alignment mechanism to carefully align the corresponding regions of a moving person and adopt an effective feature aggregation strategy to explore accurate spatiotemporal cues. STA utilizes attention to aggregate roughly-partitioned region information but suffers from slight performance degradation due to neglecting the misalignment problem. The previous best results are reported by CSACSE and SCAN, which incorporate optical flow to represent the motion information. However, off-line optical flow extraction is time and resources consuming due to the storage requirements, especially for the large-scale dataset, such as MARS, and the whole network can not be trained end to end. Given that our model is trained in an end-to-end manner, we abandon additional optical flow extraction and significantly surpass the existing works that do not use optical flow in terms of the top-1 accuracy by **6.7%**, **2.3%** and **0.7%** and the mAP by **4.5%** on three datasets respectively.

Ablation Study

To demonstrate the effectiveness of each component of our model, we compared our full model with several different settings on iLIDS-VID, PRID2011 and MARS. The overall

Table 1: Comparisons of our proposed method with the state-of-the-art on iLIDS-VID, PRID2011 and MARS datasets. Top-1, -5, -10, -20 accuracy (%) and mAP (%) are reported. '**OF**' denotes optical flow.

Method	iLIDS-VID				PRID2011				MARS			
	Top1	Top5	Top10	Top20	Top1	Top5	Top10	Top20	Top1	Top5	Top20	mAP
STFV3D (Liu et al. 2015)	44.3	71.7	83.7	91.7	64.7	87.3	89.9	92.0	-	-	-	-
TDL (You et al. 2016)	56.3	87.6	95.6	98.3	56.7	80.0	87.6	93.6	-	-	-	-
RNN (McLaughlin, Rincon, and Miller 2016)	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0	-	-	-	-
CNN+XQDA (Zheng et al. 2016)	53.0	81.4	-	95.1	77.3	93.5	-	99.3	65.3	82.0	89.0	47.6
SeeForest (Zhou et al. 2017)	55.2	86.5	-	97.0	79.4	94.4	-	99.3	70.6	90.0	97.6	50.7
ASTPN (Xu et al. 2017)	62.0	86.0	94.0	98.0	77.0	95.0	99.0	99.0	44.0	70.0	81.0	-
QAN (Liu, Yan, and Ouyang 2017)	68.0	86.8	95.4	97.4	90.3	98.2	99.3	100.0	73.7	84.9	91.6	51.7
RQEN (Song et al. 2018)	77.1	93.2	97.7	99.4	91.8	98.4	99.3	99.8	77.8	88.8	94.3	71.7
STAN (Li et al. 2018b)	80.2	-	-	-	93.2	-	-	-	82.3	-	-	65.8
M3D (Li, Zhang, and Huang 2019)	74.0	94.33	-	-	94.4	100.0	-	-	84.4	93.8	97.7	74.0
CSACSE (Chen et al. 2018)	79.8	91.8	-	-	88.6	99.1	-	-	81.2	92.1	-	69.4
SCAN (Zhang et al. 2019)	81.3	93.3	96.0	98.0	92.0	98.0	100.0	100.0	86.6	94.8	97.1	76.7
STA (Fu et al. 2019)	-	-	-	-	-	-	-	-	86.3	95.7	98.1	80.8
RRU (Liu et al. 2019)	84.3	**96.8**	-	99.5	92.7	98.8	-	98.8	84.4	93.2	96.3	72.7
CSACSE+**OF** (Chen et al. 2018)	85.4	96.7	**98.8**	99.5	93.0	99.3	100.0	100.0	86.3	94.7	**98.2**	76.1
SCAN+**OF** (Zhang et al. 2019)	**88.0**	96.7	98.0	**100.0**	95.3	99.0	100.0	100.0	87.2	95.2	98.1	77.2
FGRA(Ours)	**88.0**	96.7	98.0	99.3	**95.5**	**100.0**	**100.0**	**100.0**	**87.3**	**96.0**	98.1	**81.2**

Table 2: Component analysis of the proposed method. Top-1, -5 accuracy (%) and mAP (%) are reported. AVG, GB, RAM, TA, TAS, CR and SA indicate average pooling, global branch, region alignment mechanism, temporal attention, temporal alignment score, consistency regularization and spatial attention, respectively.

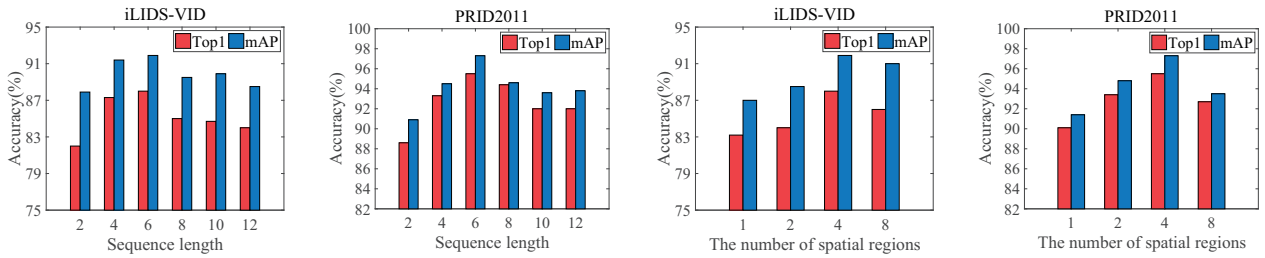
Method	iLIDS-VID			PRID2011			MARS		
	Top1	Top5	mAP	Top1	Top5	mAP	Top1	Top5	mAP
1. AVG	76.8	91.3	81.4	86.1	96.2	87.0	76.6	88.0	72.0
2. GB	55.1	80.8	60.0	73.4	89.0	75.7	58.0	79.7	46.3
3. GB+AVG	80.3	93.3	84.3	88.8	96.8	90.8	81.7	93.8	74.7
4. GB+RAM+AVG	83.3	96.0	88.5	92.3	98.4	92.4	84.6	94.6	79.0
5. GB+RAM+TA	85.3	**97.3**	90.2	94.0	99.3	94.6	86.3	95.2	80.2
6. GB+RAM+TAS	86.0	96.7	90.8	93.4	99.0	94.5	86.0	94.8	79.9
7. GB+RAM+TAS+CR	87.3	96.7	91.4	95.3	100.0	95.7	86.7	95.1	80.9
8. GB+RAM+TAS+CR+SA	**88.0**	96.7	**91.9**	**95.5**	**100.0**	**97.3**	**87.3**	**96.0**	**81.2**

results are shown in Table 2. We also conduct the hyperparameter analysis and exhibit the comparison results in Fig. 4.

Component Analysis of the Proposed Model. We specify the first variant (1) as our baseline model. It horizontally partitions the backbone feature maps into several region features and conducts temporal average pooling to calculate the region video representations. Finally, the matching feature is produced by concatenating these independent region representations. The comparison results of variants (1) (2) and (3) show that the global feature can provide complementary information for discrimination and compensate for the weak capability of matching. Given that misalignment is a vital factor that constrains the learning of high-quality matching feature, to address this, the region alignment mechanism is embedded to extract the region-aligned features before feature aggregation. It can be observed that variant (4) boost the top-1 accuracy by 3.0%, 3.5% and 2.9% and the mAP by 4.2%, 1.6% and 4.3% on three datasets. We further visualize the response maps of the region-aligned features in Fig. 5. As we can see, each response map in the same row has consistently high saliency on the same region boxed in the reference image, which confirms the effectiveness of the proposed alignment mechanism. Compared with STAN which relied on attention mechanism to learn to roughly identify and localize regions of interests for spatial alignment, we similarly introduce temporal attention module in the local branch for feature aggregation in variant (5). As we can see, the performance gap between variant (5) and STAN reflects the superiority of the proposed alignment mechanism.

One typical way to aggregate temporal information is to learn temporal attention, which requires the deployment of additional convolution layers. To simplify the network, we introduce the temporal alignment scores for temporal aggregation in variant (6) but also achieve competitive performance. In variant (7), the consistency regularization term is incorporated to guarantee the intra-consistency among the corresponding regions, which further improves the performance on all evaluation protocols with a margin. In order to verify the necessity of mining cross-region and spatial information for the prominent capability of matching, we leverage spatial attention to further capture latent spatial information and yield our full model. The best performance shown in Table 2 validates our assumption.

Sequences with Different Lengths. To investigate how



(a) (b) (c) (d)

Figure 4: Parameter analysis of the sequence length and the number of spatial regions on iLIDS-VID and PRID2011datasets. The top-1 accuracy (%) and mAP (%) are reported.

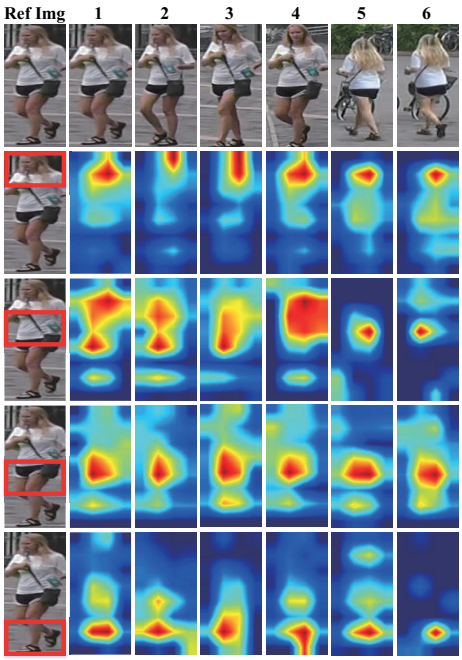


Figure 5: Examples of feature response maps extracted from sampled frames in a video. The leftmost column exhibits the reference frame and the partitioned regions boxed by red rectangles. The feature response maps deployed in the same row represent the corresponding aligned receptive fields.

the sequence length impacts the final performance, we conduct a series of experiments using various lengths of the input tracklets on two candidate datasets in Fig. 4. It can be observed that, except for the sequence length of 2, other settings can consistently achieve competitive performance, and the top-1 accuracy and mAP steadily fluctuate within 5%, reflecting that the trained model is robust to the sequence length. In our model, the sequence length of 6 is chosen for its best performance.

Various Numbers of Spatial Regions. We also investigate the performance of the proposed model using different numbers of spatial regions in Fig. 4. Considering the actual spatial size of the backbone feature map, we conduct experiments with four numbers of spatial regions: 1, 2, 4 and 8. When the number of spatial regions is set as 1, that is to say, the model tends to globally align the whole body. As the number of spatial regions increases, there is a general improvement in performance, which implies that multiple spatial regions can capture more fine-grained visual information. Our model achieves the best results with the setting of 4 spatial regions and the accuracy drops when the number of spatial regions increases from 4 to 8. This is because very small regions are easily corrupted by noise and make the overall training process unstable.

Conclusion

In this paper, we propose a frame-guided region-aligned model (FGRA) to solve the video-based person re-identification problem. To tackle the misalignment problem, a novel region alignment mechanism and an effective feature aggregation strategy are proposed to achieve region alignment and explore accurate spatiotemporal cues. Concretely, we present a frame-guided feature learning strategy and a non-parametric alignment module that uses depth-wise cross correlation to align body regions across frames, which are supervised by a consistency regularization term to improve the compactness of the region-aligned features. To form the final representation, we design temporal alignment scores and spatial attention to aggregate discriminative region information in the temporal and spatial dimension, respectively. The outstanding performance on three mainstream datasets demonstrates the effectiveness of each component of our model and the superiority of the proposed method to the existing video-based person re-identification methods.

Acknowledgements

The work is supported by National Key R&D Program of China (2018YFC0309400), National Natural Science Foundation of China (61871188), Guangzhou city science and technology research projects(201902020008).

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *ECCV*, 850–865. Amsterdam, The Netherlands: Springer.

Bolle, R. M.; Connell, J. H.; Pankanti, S.; Ratha, N. K.; and Senior, A. W. 2005. The relation between the roc curve and the cmc. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, 15–20. Buffalo, NY, USA: IEEE.

Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; and Van Gool, L. 2017. One-shot video object segmentation. In *CVPR*, 221–230. Amsterdam, The Netherlands: IEEE.

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*, 1169–1178. Salt Lake City, Utah, USA: IEEE.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778. Las Vegas, USA: IEEE.

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 *Scandinavian conference on Image analysis*, 91–102. Berlin, Heidelberg: Springer.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019. Vrstc: Occlusion-free video person re-identification. In *CVPR*, 7183–7192. Long Beach, CA: IEEE.

Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018a. High performance visual tracking with siamese region proposal network. In *CVPR*, 8971–8980. Salt Lake City, Utah, USA: IEEE.

Li, S.; Bak, S.; Carr, P.; and Wang, X. 2018b. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 369–378. Salt Lake City, Utah, USA: IEEE.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 4282–4291. Long Beach, CA: IEEE.

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, volume 33, 8618–8625. Honolulu, Hawaii, USA: AAAI Press.

Liu, K.; Ma, B.; Zhang, W.; and Huang, R. 2015. A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV*, 3810–3818. Santiago, Chile: IEEE.

Liu, H.; Jie, Z.; Jayashree, K.; Qi, M.; Jiang, J.; Yan, S.; and Feng, J. 2017. Video-based person re-identification with accumulative motion context. *IEEE TCSVT* 28(10):2788–2802.

Liu, Y.; Yuan, Z.; Zhou, W.; and Li, H. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, volume 33, 8786–8793. Honolulu, Hawaii, USA: AAAI Press.

Liu, Y.; Yan, J.; and Ouyang, W. 2017. Quality aware network for set to set recognition. In *CVPR*, 5790–5799. Honolulu, Hawaii, USA: IEEE.

McLaughlin, N.; Rincon, J. M. D.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 1325–1334. Las Vegas, USA: IEEE.

Perai, F.; Khoreva, A.; Benenson, R.; Schiele, B.; and Sorkine-Hornung, A. 2017. Learning video object segmentation from static images. In *CVPR*, 2663–2672. Honolulu, Hawaii, USA: IEEE.

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*. New Orleans, Louisiana, USA: AAAI Press.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*, 688–703. Zurich, Switzerland: Springer.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*, 499–515. Amsterdam, The Netherlands: Springer.

Wu, L.; Shen, C.; and Hengel, A. v. d. 2016. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*.

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*, 4733–4742. Venice, Italy: IEEE.

You, J.; Wu, A.; Li, X.; and Zheng, W.-S. 2016. Top-push video-based person re-identification. In *CVPR*, 1345–1353. Las Vegas, USA: IEEE.

Zhang, R.; Li, J.; Sun, H.; Ge, Y.; Luo, P.; Wang, X.; and Lin, L. 2019. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE TIP* 28(10):4870–4882.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*, 1116–1124. Santiago, Chile: IEEE.

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*, 868–884. Amsterdam, The Netherlands: Springer.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017a. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 1318–1327. Honolulu, Hawaii, USA: IEEE.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2017b. 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*, 4747–4756. Honolulu, Hawaii, USA: IEEE.