BiCnet-TKS: Learning Efficient Spatial-Temporal Representation for Video Person Re-Identification

Ruibing Hou[1,2], Hong Chang[1,2], Bingpeng Ma[2], Rui Huang[3], Shiguang Shan[1,2,4]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Shenzhen Institute of Artificial Intelligence and Robotics for Society,
The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China
[4]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, 200031, China
ruibing.hou@vipl.ict.ac.cn, bpma@ucas.ac.cn, ruihuang@cuhk.edu.cn, {changhong, sgshan}@ict.ac.cn

Abstract

In this paper, we present an efficient spatial-temporal representation for video person re-identification (reID). Firstly, we propose a Bilateral Complementary Network (BiCnet) for spatial complementarity modeling. Specifically, BiCnet contains two branches. Detail Branch processes frames at original resolution to preserve the detailed visual clues, and Context Branch with a down-sampling strategy is employed to capture long-range contexts. On each branch, BiCnet appends multiple parallel and diverse attention modules to discover divergent body parts for consecutive frames, so as to obtain an integral characteristic of target identity. Furthermore, a Temporal Kernel Selection (TKS) block is designed to capture short-term as well as long-term temporal relations by an adaptive mode. TKS can be inserted into BiCnet at any depth to construct BiCnet-TKS for spatial-temporal modeling. Experimental results on multiple benchmarks show that BiCnet-TKS outperforms state-of-the-arts with about 50% less computations. The source code is available at https://github.com/blue-blue272/BiCnet-TKS.

1. Introduction

Person re-identification (reID) [34, 50, 11] aims at retrieving a particular person across multiple non-overlapped cameras. Recently, with the emergence of large video benchmarks [50, 20] and the growth of computational resource, video person reID has been attracting a lot of attention. The video data contain richer spatial and temporal clues, which can be utilized to reduce visual ambiguities for more robust reID.

Despite the significant progress in video reID, most existing methods do not take full advantage of the rich spatial-temporal clues in videos. For *spatial clues*, most methods [28, 26, 12] conduct the same operation on each frame at same input resolution, resulting in highly redundant spatial features for consecutive frames. The redundant features easily focus on the same most representative local region [14], which may be indistinguishable for the two persons with seemingly similar local body parts. For example, as shown in Fig. 1 (a), the green T-shirt of the sequence pair attracts the most attention, but is difficult to distinguish the two pedestrians. Therefore, it is desirable to automatically capture the diverse spatial clues across consecutive frames to form a full characteristic of each identify.

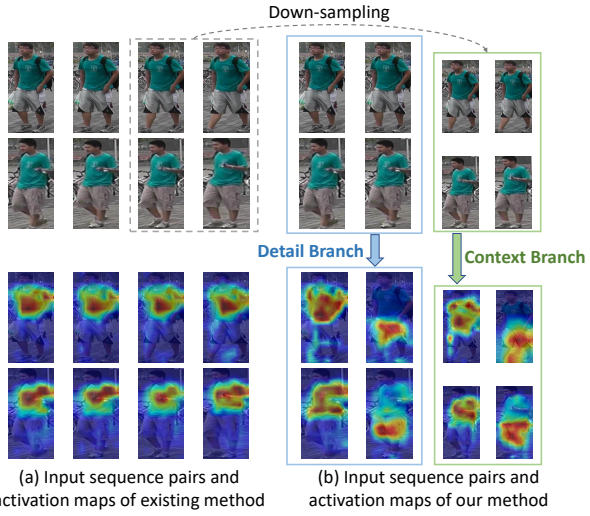


Figure 1: An example of class activation maps [53] of a pair of input video sequences of existing method [50] and our method.

For *temporal clues*, most existing methods only model

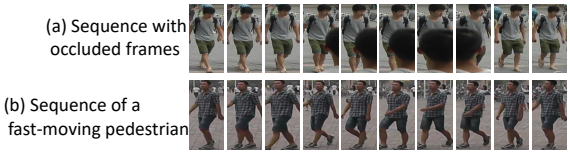


Figure 2: Short and long-term temporal relations have varying importance for different sequences. (a) A sequence with partial occlusion. The long-term temporal clues are desired to alleviate occlusion. (b) A sequence of a fast-moving pedestrian. The short-term temporal clues are desired to model detailed motion patterns.

either short-term [28, 44, 9] or long-term temporal relations [39, 46, 13]. To enhance the temporal modeling ability, a few works [20, 21] attempt to jointly capture short and long-term temporal relations and fuse the two relations with equal weights. However, the two temporal relations have varying importance for different sequences. For example, as shown in Fig. 2, for a sequence with partial occlusion, the long-term temporal relations are more important to alleviate occlusion. For a fast-moving pedestrian sequence, the short-term temporal relations play a greater role to model the detailed motion patterns. So it is necessary to *adaptively* capture short and long-term temporal relations of videos.

To explicitly fulfill above goals, we present an efficient spatial-temporal representation for video reID. We first propose a *Bilateral Complementary Network* (BiCnet) to extract complementary spatial features across consecutive frames. **Firstly**, BiCnet contains two scale-specific branches, *Detail Branch* operating on frames at original resolution to retain spatial details, and *Context Branch* processing frames at down-sampled resolution to enlarge receptive field for long-range contexts. As shown in Fig. 1 (b), with larger receptive field, the third-frame feature of the first sequence can capture broader visual clues of a green T-shirt with a backpack strap on it, which can help differentiate the two similar pedestrians. **Then** on each branch, BiCnet appends multiple parallel spatial attention modules. By enforcing the diversity of individual attention modules, the attention modules can focus on different regions for consecutive frames. As shown in Fig. 1 (b), with the diverse attention modules, the consecutive-frame features from same branch can focus on complementary body regions, covering the whole body of the target identity. **Finally**, BiCnet aggregates the complementary features from the two branches to a comprehensive spatial representation.

Furthermore, we develop a *Temporal Kernel Selection* (TKS) block to *adaptively* model the short and long-term temporal relations. Utilizing both small kernel and large kernel along the temporal dimension can capture the short and long-term temporal relations simultaneously. So TKS is designed to contain several parallel temporal convolution paths with various kernel sizes. More importantly, TKS selects a dominant temporal scale according to the global information from the multiple paths. With the selection strategy, TKS can adaptively vary the scale of temporal modeling depending on the properties of input videos, thereby exhibiting stronger temporal representational capability. TKS is computationally lightweight and imposes a slight increase in model complexity. It can be readily inserted into BiCnet, called "BiCnet-TKS", to progressively learn spatial-temporal patterns.

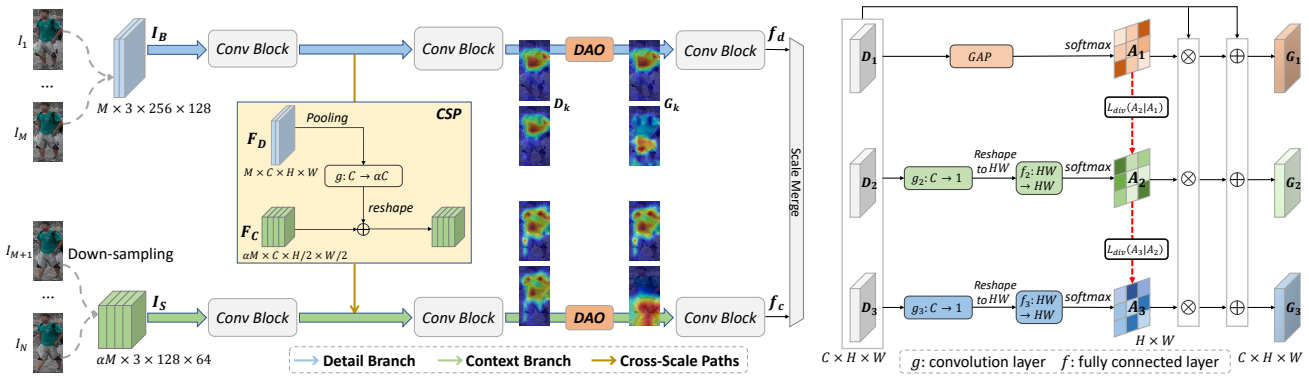
We evaluate our approach on multiple challenging video reID benchmarks. The evaluations show that our approach outperforms state-of-the-arts. Moreover, by down-sampling some frames to low-resolution, BiCnet-TKS greatly reduces the computations, requiring about 50% less computation cost than state-of-the-arts.

2. Related Work

Person ReID. Existing video reID methods mainly focus on exploiting rich spatial-temporal clues in videos. For spatial clues, most works [50, 43, 26, 49] apply temporal average pooling or a weighting strategy to fuse frame features. For temporal clues, existing methods use optical flow [28, 54, 43], recurrent neural network [28, 44, 37, 3], 3D convolution [24, 9] or non-local block [39, 12, 13] to model the temporal relations. Recently, the works [21, 20] propose to jointly capture short and long-term temporal relations. However, these methods fuse the two temporal relations with equal weights. In contrast, our TKS adaptively selects a dominate temporal relation based on the input video, exhibiting stronger temporal modeling capability.

The most similar work to ours BiCnet is TCLNet [14], which also extracts complementary features for consecutive frames. BiCnet has several advantages over it. First, TCLNet only considers one spatial scale to focus on local details, while our method is built on a two-branch architecture, which can capture both detailed features as well as long-range contexts. Second, TCLNet uses *hard erasing* to drop the salient features which may deteriorate the representation capacity, while our method adopts *soft attention* to flexibly determine the regions that should be attended to. Third, TCLNet uses multiple expensive CNNs to mine diverse parts. Our method uses diverse and lightweight attention modules with sharing CNNs, which is more computational efficient and parametric friendly.

Multi-branch Architecture. Multi-branch architecture has exhibited great success in image based vision tasks. For example, M3DNet [17] and HR-Nets [33] propose the networks that contain multiple branches and each branch has it own spatial resolution, respectively for image classification and pose estimation. The works [5, 25] propose a pyramidal feature learning network that consists of multiple scale-specific feature learning branches for image reID. However, above methods process each image at multiple resolutions, incurring additional computations. On the con-



(a) Bilateral Complementary Network (BiCnet)

(b) Diverse Attentions Operation (DAO) in BiCnet

Figure 3: (a) The overall framework of BiCnet. BiCnet contains two branches. *Detail Branch* processes frames at original resolution to encode detailed spatial clues, and *Context Branch* processes frames at half of original resolution to provide larger receptive field for long-range contexts. The input frames of a sequence are split into different branches. *Cross-Scale Paths* (CSP) fuse the two branches after each stage. *Diverse Attentions Operation* (DAO) is added on each branch which enforces consecutive frames to focus on different body regions, so as to obtain an integral characteristic of each identity. (b) The structure of DAO where a three-frame case is shown, and the divergence term $L_{div}(A_3|A_1)$ is omitted for clarity.

trary, our approach uses an individual resolution for each frame which largely reduces the computational cost. Moreover, very few methods explore the multi-branch architecture for efficient video understanding. SlowFast Networks [8] rely on a similar two-branch structure, but each branch encodes different frame rates, while our method processes frames with different spatial resolutions.

Attention Model. Attention mechanism has proven to be a potential way to enhance CNNs. SENet [16] proposes an efficient channel attention module. CBAM [40] and BAM [29] further introduce spatial attention block. SKNet [23] brings the feature attention across two spatial convolutions. Recent methods [38, 7, 47, 15] further improve the channel attention block. However existing methods are usually designed to enhance spatial representational capability. In contrast, our TKS adopts attention over different temporal kernels, which can boost the temporal representational power of video networks. Also, our BiCnet is the first work to use *diverse attention modules* across *consecutive frames* to enhance the video representation.

3. Our Approach

We aim at developing an efficient spatial-temporal representation for video reID. Our method includes two novel components, *i.e.*, BiCnet for complementary spatial representations across consecutive frames, and TKS for adaptively modeling the short and long-term temporal relations.

3.1. Bilateral Complementary Network

As shown in Fig. 1 (a), most existing methods extract highly redundant features for consecutive frames that only highlight a local body part [14]. To this end, we design a Bilateral Complementary Network to mine complementary visual clues from consecutive frames. As shown in Fig. 3, BiCnet is built on a two-branch architecture and adds a Diverse Attentions Operation (DAO) on each branch. The two-branch architecture is used to model complementary scales for different video sub-segments, and DAO is utilized to mine complementary body parts for consecutive frames. By adding DAO on each branch, BiCnet can obtain an integral characteristic of the target person, producing a comprehensive spatial representation.

Two-branch Architecture. As shown in Fig. 3 (a), BiCnet contains two CNN branches, a Detail Branch processing former several frames of given video segment at original resolution and a Context Branch operating on remaining frames at half of original resolution. By downsampling input frames to small size, Context Branch provides larger receptive field to encode long-range spatial contexts, which can complement the detailed features extracted by Detail Branch. Concretely, suppose a video segment $I = \{I_n\}_{n=1}^{N}$ contains N consecutive frames and n is the index of the video frame. We firstly divide I into two subsegments, namely big frames $I_B = \{I_n\}_{n=1}^{M}$ at original resolution, and small frames $I_S = \{I_n\}_{n=M+1}^{N}$ at half of the original resolution, where M is a hyper-parameter that determines the ratio of the small frames to big frames α. Then I_B and I_S are fed into Detail Branch (CNN_D) and Context Branch (CNN_C) separately, to obtain the corresponding feature vectors f_d and f_c as follows,

$$
\begin{aligned}
f_d &= \frac{1+\alpha}{N} \sum_k \text{CNN}_\text{D}(I_k), I_k \in \{I_n\}_{n=1}^{\frac{N}{1+\alpha}} \\
f_c &= \frac{1+\alpha}{\alpha N} \sum_k \text{CNN}_\text{C}(I_k), I_k \in \{I_n\}_{n=\frac{N}{1+\alpha}+1}^{N}.
\end{aligned} \tag{1}
$$

Finally, we simply average f_d and f_c to obtain the video feature for recognizing.

Cross-Scale Paths. Further, we add Cross-Scale Paths (CSP) that propagate the intermediate information of Detail Branch to Context Branch. CSP enables Context Branch to aware the features extracted by Detail Branch, such that Context Branch can focus on exploiting long-range visual clues less activated by the other branch.

The structure of CSP is illustrated in Fig. 3 (a). Formally, let $F_D \in \mathbb{R}^{M\times C\times H\times W}$ and $F_C \in \mathbb{R}^{\alpha M\times C\times \frac{H}{2}\times \frac{W}{2}}$ be the intermediate video feature map extracted by the same stage of Detail Branch and Context Branch respectively, where C, H and W denote the number of channels, the height and the width of feature map of big frames respectively. F_D and F_C have different spatial and temporal dimensions, so CSP first performs transformation on F_D to $\overline{F_D} \in \mathbb{R}^{\alpha M\times C\times \frac{H}{2}\times \frac{W}{2}}$ to match the size as:

$$\overline{F_D} = \mathcal{R}\left(W_c * \mathcal{P}(F_D)\right). \tag{2}$$

Here $\mathcal{P}$ is the pooling operation that performs max pooling with stride 2 to match the spatial dimension, $*$ is the convolution operation, $W_c \in \mathbb{R}^{1\times 1\times C\times \alpha C}$ is the parameter of the convolution operation, and $\mathcal{R}$ is the reshape operation reshaping the convolutional result with size $M\times \alpha C\times \frac{H}{2}\times \frac{W}{2}$ to $\alpha M \times C \times \frac{H}{2} \times \frac{W}{2}$ to math the temporal dimension. At last, $\overline{F_D}$ is fused into F_C by element-wise summation.

Diverse Attentions Operation. As shown in Fig. 3 (a), although the big frames and small frames can provide some complementary clues (*e.g.*, detailed T-shirt/additional long-distance knapsack strap feature), the frames on each branch still easily focus on around the most representational region (*e.g.*, upper-clothes). To this end, we design Diverse Attentions Operation to mine complementary regions for consecutive frames. By adding DAO on each branch, BiCnet can discover abundant discriminative parts and produce an integral complementary characteristic of each identity.

As shown in Fig. 3 (b), DAO contains several parallel attention modules and uses a specific attention module for each frame. By encouraging diversity among the generated attention maps, the attention modules can attend to complementary parts, so as to acquire diverse discriminative features for the consecutive frames.

In particular, DAO takes F_D (or F_C) as input, and uses a specific attention module for each frame feature map $(F_D)_k \in \mathbb{R}^{C\times H\times W}$. We take F_D as an example, and denote $(F_D)_k$ as D_k for simplicity. Firstly, as pointed out by [6], the intensity of each pixel in high-level feature map is proportional to the discriminative power. So we compress D_1 by channel-wise average pooling to locate the region activated by D_1, producing a self-attention map $A_1 \in \mathbb{R}^{H\times W}$:

$$A_k = \textit{softmax}\left(\frac{1}{C}\sum_{c=1}^{C}(D_k)_c\right), \quad k = 1, \tag{3}$$

Then we introduce parallel attention modules to learn to mine different and non-activated regions. Specifically, given D_k (k>1), the corresponding attention module first takes a convolutional layer to compress the channel dimension and reshapes the result to $\mathbb{R}^{HW}$. After that a fully-connected layer is applied to embed the global spatial contexts. Finally, the result is reshaped to $\mathbb{R}^{H\times W}$ followed by a softmax layer to produce corresponding attention map $A_k \in \mathbb{R}^{H\times W}$ (k>1).

In order to guide different attention modules to activate diverse regions, the corresponding spatial attention maps should be different. To achieve this, a *divergence regularization term* is introduced to measure the diversity of two attention maps A_k and A_l, which is defined as:

$$L_{div}(A_k|A_l) = 1 - \textit{sim}(A_k, A_l), \tag{4}$$

where $\textit{sim}(A_k, A_l)$ computes the similarity of A_k and A_l. Any distance measure is applicable, and we use the dot-product similarity [39] since dot-product is more implementation-friendly in modern deep learning platforms. Then the divergence loss is calculated as:

$$L = \frac{-1}{M-1}\sum_{k=2}^{M}\left(\frac{1}{k-1}\sum_{l=1}^{k-1}L_{div}\left(A_k|A_l\right)\right). \tag{5}$$

L is used to guide the optimization of parallel attention modules. When any two attention modules focus on similar person region, the generated attention maps would have a low diversity value, producing a high loss value L. So optimizing with L can drive the different attention modules to focus on different person regions. Next, we encode the diverse attention information into input feature maps by a residual operation.

At last, the updated feature maps are fed into the subsequent convolutional layers to generate feature vectors embedded with complementary visual clues.

3.2. Temporal Kernel Selection Block

Following [30, 42], we factor the video network to treat spatial clues and temporal relations separately. With the efficient BiCnet to fully mine the spatial clues, we build a *Temporal Kernel Selection* block to jointly model the short-term and long-term temporal relations. Since the temporal relations with different scales have varying importance for different sequences (as illustrated in Fig. 2), TKS combines the multi-scale temporal relations in a dynamic way, *i.e.*, different weights are assigned to different temporal scales according to input sequences.

In particular, TKS takes a sequence of consecutive-frame feature maps $F = \{F_t\}_{t=1}^{T}$ as input, where F_t is the feature map of the t^{th} frame, and conducts a triple of operations, *Partition*, *Select* and *Excite* on F.

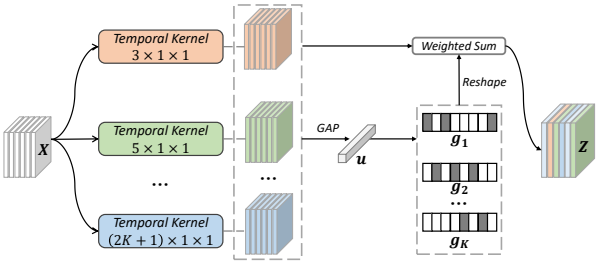


Figure 4: The architecture of Temporal Kernel Selection block.

***Partition* Operation.** Due to imperfect person detection algorithm, the adjacent frames of a video are not well aligned, which might make the temporal convolution ineffective on video reID [9]. Following [34], we use the partition strategy to alleviate the spatial misalignment issue. Specifically, given video feature map $\{F_t\}_{t=1}^{T}$, we divide each frame feature map into $h \times w$ spatial regions uniformly, and perform average pooling on each divided region to construct a region-level video feature map $X \in \mathbb{R}^{T \times C \times h \times w}$.

***Select* Operation.** As shown in Figure. 4, given X, we conduct K parallel paths $\{\mathcal{F}^{(i)} : X \rightarrow Y^{(i)} \in \mathbb{R}^{T \times C \times h \times w}\}_{i=1}^{K}$, where $\mathcal{F}^{(i)}$ is 1D temporal convolution [30] with kernel size $2i+1$. For further efficiency, the temporal convolution with a $(2i+1) \times 1 \times 1$ kernel is replaced with dilated convolution with a $3 \times 1 \times 1$ kernel and dilation size i. The basic idea of *select* operation is to use global information from all temporal paths to determine the assigned weights to each path. In particular, we first fuse the outputs of all paths by element-wise summation, then perform global average pooling to obtain a global feature $u \in \mathbb{R}^{C \times 1}$:

$$u = GAP_{T,h,w}(\sum_{i=1}^{K} Y^{(i)}), \tag{6}$$

where $GAP_{T,h,w}$ denotes global average pooling along the temporal and spatial dimension. After that the channel selection weights $\{g_i \in \mathbb{R}^{C \times 1}\}_{i=1}^{K}$ are obtained according to the global embedding u,

$$g_i = \frac{\exp\left(W_i u\right)}{\sum_{j=1}^{K} \exp\left(W_j u\right)} \quad i \in \{1, \ldots, K\}, \tag{7}$$

where $W_i \in \mathbb{R}^{C \times C}$ is the transformed parameters to generate g_i for $Y^{(i)}$. The aggregated feature map $Z \in \mathbb{R}^{T \times C \times h \times w}$ is then obtained through the selection weights on various temporal kernels,

$$Z = \sum_{i=1}^{K} \mathcal{R}(g_i) \odot Y^{(i)}, \tag{8}$$

where $\mathcal{R}$ is the reshape operation reshaping $g_i \in \mathbb{R}^{C \times 1}$ to $\mathbb{R}^{1 \times C \times 1 \times 1}$ to be compatible with the size of $Y^{(i)}$.

It is worth pointing out that, in contrast to using scale-wise weight to provide coarse fusion, we choose to use channel-wise weights (Eq. 7) for fusing. This design results in more fine-grained fusion that tunes each feature channel. In addition, the weights are dynamically computed conditioned on input videos. This is crucial for reID where different sequences may have different dominate temporal scales.

***Excite* Operation.** The *excite* operation modulates the input feature map by conditioning on Z with a residual scheme. The final feature map $E = \{E_t\}_{t=1}^{T}$ is obtained as: $E_t = \mathcal{U}(Z_t) + F_t$. Here $\mathcal{U}$ is the nearest neighbor upsampler that performs upsampling on Z_t to match the spatial resolution of F_t. TKS block maintains the input size, thus can be inserted at any depth of BiCnet to extract efficient spatial-temporal feature.

3.3. Overall Architecture

Our idea of BiCnet is generic, and it can be instantiated with different backbones [36, 35, 10]. Following recent works [9, 20, 32], we use ResNet-50 [10] pretrained on ImageNet [19] with last down-sampling operation removed as the backbone. The branches of BiCnet are built on ResNet-50 that consists of four consecutive stages, *i.e.*, $stage_1 \sim stage_4$. Diverse Attentions Operation is added after $stage_3$ since the high-level feature maps contain more semantic information. TKS block can be inserted into BiCnet to any stage to construct BiCnet-TKS for spatial-temporal modeling.

Structure and Weight Sharing between Branches. An immediate problem of multi-branch architecture [5] is that it introduces several times parameters and incurs a higher risk of overfitting. So we use the same structure and share the parameters for the two branches of BiCnet. It reduces the number of parameters and makes BiCnet need no extra parameters over single-branch reID network.

Computation Cost Analysis. To illustrate the computation cost of BiCnet-TKS, we consider a common video reID Baseline [50] that uses ResNet-50 to extract feature for each frame at original resolution. Assume that the FLOPs for Baseline to extract one-frame feature is p, Baseline requires Np FLOPs to process a video with N frames. BiCnet-TKS splits the video frames to big frames at original resolution and small frames at half of original resolution by a ratio $1:\alpha$ (Eq. 1). So BiCnet-TKS requires about $\frac{N}{1+\alpha}p + (\frac{\alpha N}{1+\alpha})\frac{p}{4}$ FLOPs[1], corresponding to about $\frac{3}{4} - \frac{3}{4\alpha+4}$ relative decrease over Baseline.

We can see that the the computation cost decreases as α increases. However, when α is too large, the small frames would dominate the network optimization, causing a severe performance drop. We experimentally observe that setting α to 3 offers the best trade-off between computation

[1]The computations of CSP, DAO and TKS are negligible compared to the feature extraction of ResNet-50.

Table 1: Comparison with state-of-the-arts on MARS, DukeMTMC-VideoReID and LS-VID datasets. The methods are separated into three groups, mainly for spatial (**S**), temporal (**T**) and spatial-temporal (**ST**) modeling.

Methods		MARS		Duke-Video		LS-VID	
		mAP	top-1	mAP	top-1	mAP	top-1
S	COSAM* [32]	79.9	84.9	94.1	95.4	-	-
	MGRAFA [48]	85.9	88.8	-	-	-	-
T	Two-stream [31]	-	-	-	-	32.1	48.2
	STMP [27]	72.7	84.4	-	-	39.1	56.8
	M3D [21]	74.1	84.4	-	-	40.1	57.7
	GLTP [20]	78.5	87.0	93.7	96.3	44.3	63.1
ST	DRSA [22]	65.8	82.3	-	-	37.8	55.8
	VRSTC [12]	82.3	88.5	93.5	95.0	-	-
	I3D [2]	83.0	88.6	-	-	33.9	51.0
	P3D [30]	83.2	88.9	-	-	35.0	53.4
	STGCN [46]	83.7	89.9	95.7	**97.3**	-	-
	IAUnet [13]	85.0	**90.2**	96.1	96.9	-	-
	TCLNet [14]	85.1	89.8	**96.2**	96.9	70.3	81.5
	AP3D [9]	85.1	90.1	95.6	96.3	73.2	84.5
	MGH [45]	85.8	90.0	-	-	-	-
ST	BiCnet-TKS	**86.0**	**90.2**	96.1	96.3	**75.1**	**84.6**

cost and accuracy. In this case, BiCnet-TKS only requires $\sim 44\%$ computation costs over Baseline, which is more efficient to extract the spatial-temporal feature.

4. Experiment

4.1. Dataset and Settings

Datasets. We evaluate the proposed method on multiple video reID datasets, *i.e.*, MARS [50], DukeMTMC-VideoReID [41] and LS-VID [20].

Evaluation Metric. We adopt mean Average Precision (mAP) [51] and Cumulative Matching Characteristics (CMC) [1] as evaluation metrics.

Implementation Details. During training, for each video sequence, we randomly sample 8 frames with a stride of four frames to form a video segment. Each batch contains 16 persons, each person with 4 video segments. We resize the split big frames to 256×128 and small frames to 128×64. The horizontal flip and random erasing [52] are adopted for data augmentation. As for the optimizer, Adam [18] with weight decay 0.0005 is adopted to update the parameters. We train the model for 150 epochs in total. The learning rate is initialized to 3.5×10^{-4} with a decay factor 0.1 at every 40 epochs. In BiCnet, the ratio of small frames to big frames is set to 3. In TKS, the number of temporal kernels is set to 2, and the divided regions is 4×2.

During testing, for each video sequence, we first split it into several 8-frame video segments. Then we extract the feature for each video segment by BiCnet-TKS and the final video feature is the averaged representation of all segments. After feature extraction, the cosine distances between the query and gallery features are computed for retrieval.

4.2. Comparison with State-of-the-arts

In Tab. 1, we compare our method with state-of-the-arts on MARS and DukeMTMC-VideoReID and LS-VID datasets. Our method achieves the best performance. It is noted that: **(1)** The spatial-based methods [32, 4, 48] process each frame by same operation and resolution, so they do not fully consider the spatial redundancy between frames. On the contrary, our BiCnet ensures different frames to focus on divergent regions to form an integral person representation and achieves better performance. **(2)** Our method outperforms TCLNet [14], with an improvement up to 4.8% mAP on LS-VID dataset. The significant improvements can be attributed to the use of two-branch architecture and flexible soft attention modules. **(3)** The temporal-based methods [3, 24, 9] lack the ability of modeling both short and long-term temporal relations. Our method outperforms these methods with an 1% mAP improvement on MARS. **(4)**. The methods [20, 21, 45] aggregate the multi-scale temporal relations with equal weights. Our method achieves better performance by an adaptive selection mechanism. **(5)**. All existing methods add computations over Baseline. In contrast, our method greatly reduces the computation cost by processing some frames at low-resolution. Overall, our method outperforms state-of-the-arts with about 50% computation budgets.

4.3. Ablation Study

In this section, we respectively investigate the effectiveness of BiCnet and TKS block by conducting a series of ablation studies on MARS dataset.

Table 2: Component Analysis of BiCnet-TKS on MARS. We also report the number of average floating-point operations (GFLOPs) for one frame, and the parameter number (Param.) of the networks.

Models	MARS			
	GFLOPs.	Param.	mAP	top-1
Base-S (128×64)	1.02	23.5M	80.7	87.4
Base-B (256×128)	4.08	23.5M	85.2	89.1
Two-branch (TB)	1.81	23.5M	84.3	89.6
TB+CSP	1.89	27.6M	**85.0**	89.6
TB+CSP+AO (wo L_1)	1.89	27.6M	85.2	89.3
TB+CSP+DAO (BiCnet)	1.89	27.6M	**85.6**	**89.8**
BiCnet-TK (fix-fusion)	1.91	29.1M	85.5	89.6
BiCnet-TKS	1.99	29.2M	**86.0**	**90.2**

Table 3: Results of Single-branch/Multi-branch Architecture with a single resolution/different combinations of multiple resolutions inputs. Height denotes the input resolution is Height×(Height/2)

Height			MARS			
256	128	64	GFLOPs.	Param.	mAP	top-1
✓			4.08	23.5M	**85.2**	89.1
	✓		1.02	23.5M	80.7	87.4
		✓	0.25	23.5M	64.1	77.4
✓	✓		2.55	23.5M	84.8	**89.4**
✓	✓	✓	1.76	23.5M	79.1	86.1

Table 4: Results of Two-branch architecture (TB) with different α (the ratio of small frames to big frames).

α	MARS			
	GFLOPs.	Param.	mAP	top-1
0 (Base-B)	4.08	23.5M	**85.2**	89.1
1	2.57	23.5M	84.8	89.4
2	2.07	23.5M	84.4	**89.7**
3	1.81	23.5M	84.3	89.6
4	1.67	23.5M	83.8	89.5
$+\infty$ (Base-S)	1.02	23.5M	80.7	87.4

4.3.1 The components of BiCnet.

To validate the effectiveness of BiCnet, we introduce a baseline that adopts ResNet-50 with temporal average pooling to generate the video feature. The baseline processes all frames at the same resolution and is trained with cross entropy and triplet loss. We consider two baseline models, *i.e.*, Base-B processing frames at original resolution (256×128), and Base-S processing frames at half of original resolution (128×64). The comparisons are shown in Tab. 2.

The influence of branch number. BiCnet is built on a two-branch architecture. It is easy to extend to multiple branches case which splits the video frames into multiple groups and uses an individual resolution for each group. In this part, we conduct an uniform split for fair comparison. The results are shown in Tab. 3. From Tab. 3, we have following observations: (1) Training ResNet-50 on frames at 128×64 resolution still offers reasonable accuracy, while saving 75% computations (measured by floating point operations). (2) Too small input resolution (64×32) causes severe performance degradation, with a drop up to 21% mAP. We argue that too small input size leads to serious loss of spatial details, which is difficult to distinguish pedestrians with small inter-class variations. (3) The three-branch architecture performs worse than two-branch structure. It is likely that the branch with 64×32 input resolution would disturb the optimization of network parameters. So we use a two-branch architecture, which can achieve comparable performance to Base-B with less computations.

Two-branch architecture *w.r.t* split ratio. We then investigate the influence of the split ratio α (in Eq. 1), *i.e*, the ratio of small frames (128×64) to big frames (256×128), to the two-branch architecture (TB). The results are shown in Tab. 4. We can observe that with α increases, TB greatly reduces the average computations of processing one frame. But the mAP of TB decreases as α increases. We argue that it is due to the lack of interaction between the two branches. In particular, the two branches of TB independently extract features, so it is difficult for one branch to learn to capture the clues ignored by the other branch. Moreover, the feature discriminative power of low-resolution frames is lower than that of high-resolution frames. So directly using low-resolution frames inevitably weakens the discrimination of final features. In addition, we observe that α=3 only brings slight drop compared to α=2. Considering computational complexity, we set α to 3 in this work.

Effectiveness of Cross-Scale Paths. We evaluate the effect of CSP by adding it after each stage of above two-branch architecture. As shown in Tab. 2, compared with TB, employing CSP brings 0.7% mAP gains with small computational overhead. We argue that with the information propagation from Detail Branch to Context Branch, Context Branch can enhance its representational power. In addition, the two branches can learn to work collaboratively to mine complementary clues, *i.e.*, Detail Branch extracts the detailed feature of local body parts, and Context Branch focuses more on the long-distance contexts, to further enhance the feature representation.

Effectiveness of Diverse Attentions Operation. Finally, we investigate the individual effect of the attention modules and divergence constraint on DAO. The results are presented in Tab. 2. The difference between TB+CSP+AO and TB+CSP+DAO is that TB+CSP+AO appends parallel attention modules without L_1 to guide optimization. As shown in Tab. 2, TB+CSP+AO achieves negligible gains over TB+CSP, which indicates that the visual features captured by different attention modules are almost the same. TB+CSP+DAO achieves 0.6% mAP improve-

Table 5: Results of BiCnet-TKS with different combinations of multiple temporal kernels in TKS.

kernel size			MARS			
K3	K5	K7	GFLOPs.	Param.	mAP	top-1
✓			1.94	28.3M	85.1	89.9
	✓		1.94	28.3M	85.3	90.1
		✓	1.94	28.3M	85.5	89.8
✓	✓		1.99	29.2M	**86.0**	**90.2**
✓		✓	1.99	29.2M	85.7	90.0
	✓	✓	1.99	29.2M	85.6	90.1
✓	✓	✓	2.04	30.0M	85.8	90.2

ment over TB+CSP, which validates the capability of the proposed divergence regularization term. We argue that the divergence loss enforces different attention modules to focus on complementary person regions and form an integral characteristic of target identity. The integral characteristic is more conductive to distinguish different identities with similar local parts.

4.3.2 The components of TKS block.

Effectiveness of TKS. We first assess the effectiveness of TKS block by adding it after $stage_2$ of BiCnet in Tab. 2. TKS brings 0.4% mAP and top-1 accuracy gains over BiCnet with an extremely small increase in computational complexity. We argue that TKS is complementary to BiCnet, *i.e.*, TKS provides the temporal features that cannot be extracted by BiCnet. Furthermore, in order to verify the effect of the adaptively selection mechanism in TKS, we introduce a *Temporal Kernel* (TK) block which simply averages the results with the multi-scale kernels ($Z = \frac{1}{K}\sum_{i=1}^{K} Y^{(i)}$ in Eq. 8). As shown in Tab. 2, TK brings no gain over BiCnet, which indicates that the improvement of BiCnet is attributed to the adaptive selection among the multi-scale kernels.

TKS *w.r.t* the number of temporal kernels (K). Next, we investigate the influence of combination of different kernels. We consider three different kernels, called "K3" (standard $3 \times 1 \times 1$ 3D convolutional kernel), "K5" ($3\times1\times1$ convolution with dilation 2 to approximate $5\times1\times1$ kernel size), and "K7" ($3 \times 1 \times 1$ convolution with dilation 3 to approximate $7 \times 1 \times 1$ kernel size). The results are shown in Tab. 5. We can observe that: (1) When using two temporal kernels with different sizes, in general the accuracy increases. The mAP and top-1 accuracy in the second block of the table ($K = 2$) are generally higher than those in the first block ($K = 1$), indicating the effectiveness of modeling both short and long-term temporal relations. (2) Using more temporal kernels ($K = 3$) does not bring performance gain, showing two temporal kernels are enough to capture the temporal clues of video.

Efficient positions to place TKS. Tab 6 compares the results of placing a TKS block to different stages of BiCnet.

Table 6: Results of BiCnet-TKS when placing TSK blocks on different stages.

Stage	MARS			
	GFLOPs.	Param.	mAP	top-1
$stage_1$	1.99	28.0M	85.3	90.1
$stage_2$	1.99	29.2M	**86.0**	90.2
$stage_3$	1.99	34.1M	85.7	**90.4**
$stage_4$	2.29	53.5M	85.4	90.0
$stage_{23}$	2.09	35.7M	85.8	90.3

It can be seen that the improvements by placing one TKS block in $stage_2$ and $stage_3$ are similar. However, placing TKS block in $stage_1$ and $stage_4$ leads to performance degradation. It is likely that the low-level features in $stage_1$ are insufficient to provide precise semantic information, thus TKS can not model temporal relations between body parts very well. And since BiCnet learns to focus on different regions for consecutive frames on $stage_3$, the frame features on $stage_4$ lack of coherent temporal relations, so TKS is not capable to extract an effective temporal feature on $stage_4$. We also observe that adding more TKS blocks does not bring gain, indicating that a TKS block is usually enough for temporal modeling.

Time Overhead. The running times are positively correlated with computation cost of models. In Tab. 2, Base-B takes 11ms to extract feature for a 8-frames sequence. While BiCnet-TKS only takes 6ms, corresponding to a 45.4% relative decrease over Base-B (both timings are performed on one NVIDIA 2080Ti GPU).

5. Conclusions

In this work, we present a computation-friendly spatial-temporal representation for video reID. Firstly, we introduce Bilateral Complementary Network. BiCnet contains two branches, Detail Branch preserving the spatial detail clues from original resolution, and Context Branch utilizing down-sampling operation to enlarge receptive field for longer-range contexts modeling. On each branch, BiCnet appends parallel and diverse attention modules to mine divergent regions for consecutive frames. Furthermore, we propose Temporal Kernel Selection block to adaptively capture temporal relations of videos. Extensive experiments demonstrate the superiority of our method over state-of-the-arts with about 50% less computations.

Acknowledgement This work is partially supported by Natural Science Foundation of China (NSFC): 61876171 and 61976203, and the Open Project Fund from Shenzhen Institute of Artificial Intelligence and Robotics for Society, under Grant No. AC01202005015 and 2019-INT006.

References

[1] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. The relation between the roc curve and the cmc. In *AUTOID*, pages 15–20, 2005.

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

[3] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018.

[4] G. Chen, Y. Rao, J. Lu, and J. Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification. In *ECCV*, pages 660–676, 2020.

[5] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *CVPR*, pages 2590–2600, 2017.

[6] J. Choe and H. Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, pages 2219–2228, 2019.

[7] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard. Attentional feature fusion. *arXiv preprint arXiv:2009.14082*, 2020.

[8] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[9] X. Gu, B. Ma, H. Chang, H. Zhang, and X. Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, pages 228–243, 2020.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770 – 778, 2016.

[11] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, pages 9317–9326, 2019.

[12] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, pages 7183–7192, 2019.

[13] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Iaunet: Global context-aware feature learning for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[14] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Temporal complementary learning for video person re-identification. In *ECCV*, pages 388–405, 2020.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *NeurIPS*, pages 9401–9411, 2018.

[16] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[17] G. Huang, D. Chen, T. Li, F. Wu, d. van, and K. Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2019.

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

[19] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[20] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, pages 3958–3967, 2019.

[21] J. Li, S. Zhang, and T. Huang. Multiscale 3d convolution network for video based person reidentification. In *AAAI*, pages 8618–8625, 2019.

[22] S. Li, S. Bak, P. Carr, C. Hetang, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018.

[23] X. Li, W. Wang, X. Hu, and J. Yang. Selective kernel networks. In *CVPR*, pages 510–519, 2019.

[24] X. Liao, L. He, and Z. Yang. Video-based person re-identification via 3d convolutional networks and non-local attention. In *ACCV*, pages 620–634, 2018.

[25] J. Liu, Z. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet cnn for person re-identification. In *ACMMM*, pages 192–196, 2016.

[26] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, pages 4694–4703, 2017.

[27] Y. Liu, Z. Yuan, W. Zhou, and H. Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, volume 33, pages 8786–8793, 2019.

[28] N. McLaughlin, J. M. del Rincon, and P. C. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016.

[29] J. Park, S. Woo, J. Lee, and I. Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[30] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5533–5541, 2017.

[31] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014.

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

[33] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019.

[34] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[36] G. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[38] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, pages 11534–11542, 2020.

[39] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[40] S. Woo, J. Park, J. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[41] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Quyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, pages 5177–5186, 2018.

[42] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

[43] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, pages 4743–4752, 2017.

[44] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716, 2016.

[45] Y. Yan, J. Qin, J. Chen, L. Liu, F. Zhu, Y. Tai, and L. Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *CVPR*, pages 2899–2908, 2020.

[46] J. Yang, W. Zheng, Q. Yang, Y. Chen, and Q. Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, pages 3289–3299, 2020.

[47] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, and R. Manmatha. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[48] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, pages 10407–10416, 2020.

[49] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X. Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR*, pages 4913–4922, 2019.

[50] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[51] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

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

[53] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[54] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 6776–6785, 2017.