Salient-to-Broad Transition for Video Person Re-identification

Shutao Bai[1,2], Bingpeng Ma[2], Hong Chang[1,2], Rui Huang[3], Xilin Chen[1,2]
[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]The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China
shutao.bai@vipl.ict.ac.cn,bpma@ucas.ac.cn,ruihuang@cuhk.edu.cn,{changhong,xlchen}@ict.ac.cn

Abstract

Due to the limited utilization of temporal relations in video re-id, the frame-level attention regions of mainstream methods are partial and highly similar. To address this problem, we propose a Salient-to-Broad Module (SBM) to enlarge the attention regions gradually. Specifically, in SBM, while the previous frames have focused on the most salient regions, the later frames tend to focus on broader regions. In this way, the additional information in broad regions can supplement salient regions, incurring more powerful video-level representations. To further improve SBM, an Integration-and-Distribution Module (IDM) is introduced to enhance frame-level representations. IDM first integrates features from the entire feature space and then distributes the integrated features to each spatial location. SBM and IDM are mutually beneficial since they enhance the representations from video-level and frame-level, respectively. Extensive experiments on four prevalent benchmarks demonstrate the effectiveness and superiority of our method. The source code is available at https://github.com/baist/SINet.

1. Introduction

In the past few years, video person re-identification (re-id) has achieved favorable progress [33, 39] with the help of CNNs [11, 19]. However, further development of video re-id remains hindered because it is challenging to effectively utilize the rich temporal information among video frames, as pointed out in [26].

Recently, some approaches [41, 42, 44] try to exploit the temporal relations for a mutual enhancement between frames. To realize such enhancement, these methods mainly adopt self-attention mechanism [38] or Graph Convolution Networks (GCNs) [5, 8] to encourage the information flow among video frames. In this way, the final frame-level features will be more rich and recognizable. Although these methods have achieved encouraging performances in video re-id, they still have several intrinsic drawbacks.

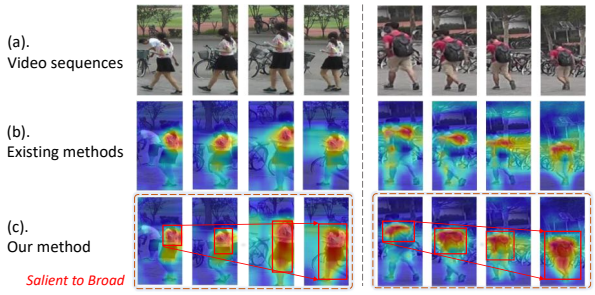

Figure 1. Comparison of existing methods and our method (SBM) with respect to Class Activation Maps (CAM) [32]. There is a clear *salient-to-broad* transition of activation maps (attention regions) in our method. Warmer color represents a higher value.

First, for each frame, the concentration of these methods is usually confined in a salient but *partial* region. In fact, when a model has focused on a partial region that can recognize a pedestrian, it will not pay attention to other regions, which results in representations with limited power. Obviously, this property should be avoided for a robust re-id model since it is desirable to use the complete characteristics of a given pedestrian. As shown in Figure 1(b), these methods pay almost all the attention to the upper clothes while ignoring the A-line skirt and other human parts. Maybe the black skirt is not as discriminative as the upper clothes in this particular example, but it is still a vital clue, especially when other pedestrians, if any, wear similar upper clothes. Therefore, enlarging the attention regions is of central importance to further enhance the robustness and discriminative ability of video embeddings.

Second, the utilization of temporal relations, *i.e.*, mutual enhancement between frames, is limited. Specifically, these methods regard the temporal relations as mutual enhancement or homogeneous information flow across all

frames. In this way, the frame-level embeddings will be richer since they contain mutually enhanced information from other frames. However, such enhancement will also mix the frame-level embeddings, making them more similar to each other, even redundant. As illustrated in Figure 1(b), all four frames focus on nearly identical regions, indicating that their embeddings are highly similar. The similarity or redundancy sacrifices the differences across frames, which limits further improvement in the final temporal fusion stage. Therefore, to make better use of temporal relations, it is desirable to leverage temporal cues from another perspective and encourage the differences between frames.

In this paper, we propose a *Salient-to-Broad Module* (SBM), which achieves the above two goals in a unified framework. SBM innovatively leverages the temporal relations to amplify the differences of frames, *i.e*., *gradually* enlarges the attention regions of consecutive frames. Specifically, we expect the pedestrians' representations to be more informative and powerful, so they should contain as much foreground information as possible. While the previous frames have focused on a salient but partial region, we require SBM to pay attention to a broader region for the later frame. In practice, SBM leverages temporal relations via *difference amplification*, which is implemented by properly broadening regions to be attended in later frames. In summary, SBM realizes the salient-to-broad transition as shown in Figure 1(c). As a result, SBM makes frame-level features more complete and diverse, thus produces more informative video-level features after temporal fusion.

Moreover, we introduce an *Integration-and-Distribution Module* (IDM) to assist our SBM. SBM increases the representation capability of video-level features by enhancing differences across frames. But the performance of SBM also depends on the richness of frame-level information. To this end, IDM will integrate and distribute the informative global features, enabling message passing across all frames. The propagation is *input-agnostic* and is constructed with all information from input data. By doing this, IDM is reciprocal to SBM: IDM consolidates the frame-level representations, and SBM will enrich the video-level representations. Thus, the combination of SBM and IDM will incur more powerful representations for video re-id.

SBM and **I**DM can be inserted into the backbone network together to form SINet. We carry out extensive experiments on four benchmarks to demonstrate the effectiveness of our method. Notably, SINet achieves *91.0%* and *87.4%* rank-1 accuracy on MARS and LS-VID, respectively, surpassing the existing state-of-the-art models.

2. Related Work

Recently, video re-id has drawn more and more attention from both academic and industrial researchers. Compared with image data, the additional temporal relations in video effectively alleviate many issues such as occlusion and motion blurs. There are tremendous methods that are designed to capture such temporal relations in different ways.

Temporal Weighted. One major stream [9, 34, 46, 49] adopts temporal attention to determine the importance of each frame, aiming to drop the low-quality frames. But these methods ignore the dependencies between frames, which hinders their further improvements.

Mutual Enhancement. Mainstream state-of-the-arts [23, 41, 42] adopt self-attention or GCNs to model temporal relations. For example, Liu *et al*. [26] use non-local block [38] to capture long-range dependencies and make each feature perceive the entire spatial-temporal space. All these methods use temporal relations for mutual enhancement. Different form these methods, we leverage the temporal relations from the perspective of difference amplification, and thus our method obtains more comprehensive and informative representations.

Others. Other methods exploit the temporal relations in many aspects including optical flow [3, 27], RNNs [30, 43], 3D CNNs [10, 24], recovery [18], and coherence constrain [4]. However, they all suffer different drawbacks such as non-global temporal modeling [10, 24], high computational cost [3, 18, 27], high-level modeling [4, 30, 43]

The recent work that is most similar to ours is TSE [17], which erases the salient features in later frames, albeit with several key fundamental differences. First, transition of attention regions in our method is *salient-to-broad* implemented by suppression, while TSE is *one-to-another* by erasing. So TSE drops the salient features in the later frames and may deteriorate the representation capability of the final embeddings. Second, the erasing region of TSE is fixed and partial, while our method can flexibly determine the suppressed region.

Moreover, capturing complete and richer visual patterns with image erasing is widely explored, especially in weakly supervised object localization [29, 40]. For example, besides the traditional localization branch, EIL [29] uses another branch that inputs the erased feature maps, and captures less discriminative regions. However, these methods are all erasing-based rather than suppression-based, *i.e*., *one-to-another*, and cannot model the interactions of the entire object. Moreover, compared with video data, these methods need to maintain another stream as EIL and thus incur more memory requirement.

3. Methodology

In this section, we will first describe SBM that makes the attention regions transit from salient towards broader ones. Then, we will elaborate IDM, which encourages information flow between consecutive frames. Finally, we will give an overview of our SINet, which is the hierarchical architecture of SBM and IDM.

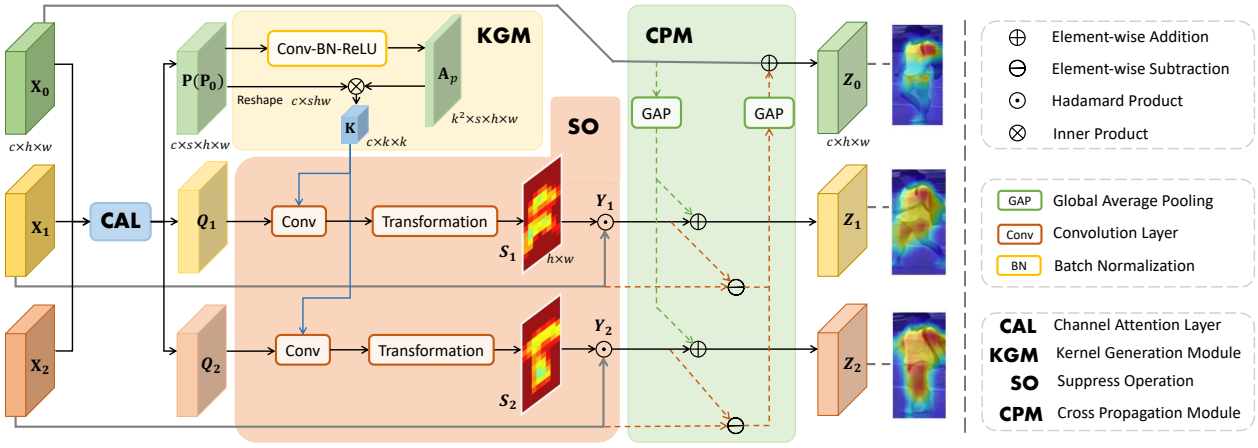

Figure 2. The detailed architecture of the Salient-to-Broad Module. For better visualization, we only describe the case when the video contains three frames. In this case, SBM takes middle-level feature maps $\mathbf{X}_{0,1,2}$ as input.

3.1. Salient-to-Broad Module

In this paper, we propose SBM to realize the salient-to-broad transition of attention regions for consecutive frames. The transition is achieved with the *difference amplification* in the temporal dimension, *i.e.*, suppressing the salient features that have been activated before. As illustrated in Figure 2, when passing through SBM, the later frames concentrate on broader regions. As a result, their frame-level embeddings will be more complete and informative, leading to powerful characteristics of pedestrians.

Notably, the additional information in the broad regions can *supplement* the salient ones. We argue that the salient features contain the most discriminative and unusual characteristics of pedestrians, which are useful to distinguish from most other identities. However, for pedestrians who also have such characteristics, the broad features are more helpful as they cover the entire foreground information. Besides, since the salient features will be activated in all frames, they still dominate discriminations while the broad features assist. Examples are illustrated in Figure 6 and 7.

The details of our SBM are elaborated as follows.

Input. The input of SBM has two terms: feature maps and split position. For feature maps, SBM adopts middle-level feature maps that have both semantic and detailed information. Specifically, given a clip $\mathbf{I} = \{I_i\}_{i=0}^{t-1}$ containing t frames, we use a backbone model to obtain the middle-lever feature maps $\{\mathbf{X}_i\}_{i=0}^{t-1}$. Here $\mathbf{X}_i \in \mathbb{R}^{c\times h\times w}$, and c, h, w are the channel size, height, and width, respectively.

As for *split position* s, it determines the split of former frames and latter frames, *i.e.*, which frames need to be suppressed. In detail, SBM will suppress the salient features in later frames $\mathbf{X}_{s,\dots,t-1}$, which have been captured in earlier frames $\mathbf{X}_{0,\dots,s-1}$. Figure 2 shows the pipeline of SBM, where $t = 3$ and $s = 1$ for conciseness.

Channel Attention Layer (CAL). The first procedure of SBM is CAL, which aims to filter out misguided and meaningless channels. Due to the zero-padding of models, some channels may focus on the periphery and tend to select the background as salient features. This may misguide the later suppression since the salient pedestrians are usually in the center of the input frames. To this end, we generate the channel weights as:

$$\mathbf{w}_{ctr} = \begin{cases} 1, & \text{GAP}\left(\mathbf{X}\right) < \text{GAP}\left(\mathbf{X}_{no\_pad}\right) \\ 0, & otherwise \end{cases} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{c\times t\times h\times w}$ is the concatenation of $\{\mathbf{X}_i\}_{i=0}^{t-1}$, and $\mathbf{X}_{no\_pad} \in \mathbb{R}^{c\times t\times (h-2)\times (w-2)}$ is the subarea of $\mathbf{X}$ without the first and last rows/columns. GAP is global average pooling. Here, $\mathbf{w}_{ctr} \in \mathbb{R}^c$ is an estimation of the centrality of each channel, *i.e.*, it will give 1 to channels that concern on the central foreground and 0 to others.

CAL also adopts a standard Squeeze-and-Excitation (SE) block [19] to selectively emphasize informative channels. The result is a c dimension vector $\mathbf{w}_{se} \in (0, 1)$, indicating the weight of each channel.

By applying $\mathbf{w}_{ctr}$ and $\mathbf{w}_{se}$ to the input feature maps, CAL will return more centralized and meaningful feature maps. For clearness, we rename the returned feature maps as $\{\mathbf{P}_i\}_{i=0}^{s-1}$ and $\{\mathbf{Q}_i\}_{i=s}^{t-1}$ based on the split position s.

$$\begin{aligned} \mathbf{P}_i &= \mathbf{w}_{ctr} \cdot \mathbf{w}_{se} \cdot \mathbf{X}_i, \quad i = 0, ..., s-1 \\ \mathbf{Q}_i &= \mathbf{w}_{ctr} \cdot \mathbf{w}_{se} \cdot \mathbf{X}_i, \quad i = s, ..., t-1. \end{aligned} \tag{2}$$

Kernel Generation Module (KGM). After CAL, we need to extract the salient features of the former frames $\{\mathbf{P}_i\}_{i=0}^{s-1}$. The extracted features will be used to suppress the salient features in the subsequent suppression procedure. Inspired by [31], SBM leverages KGM to generate

a $k \times k$ (k: kernel size) convolution kernel in order to include as many salient features of $\{\mathbf{P}_i\}_{i=0}^{s-1}$ as possible.

The overall architecture of KGM is shown in Figure 2. Specifically, the input tensor $\mathbf{P} \in \mathbb{R}^{c\times s\times h\times w}$ of KGM is the concatenation of $\{\mathbf{P}_i\}_{i=0}^{s-1}$. Then KGM adopts a multi-head spatial attention mechanism to weigh the importance of each position, and returns k^2 attention maps $\mathbf{A}_p \in \mathbb{R}^{k^2\times s\times h\times w}$.

To further increase the diversity and information richness of attention maps, KGM needs to avoid the collapse [25] of $\mathbf{A}_p$ and make these k^2 maps concentrate on different regions. To this end, we adopt L1-Normalization sequentially on the k^2 dimension and the $s\times h\times w$ dimension. After that, KGM will produce the $k \times k$ kernels $\mathbf{K} \in \mathbb{R}^{c\times k^2}$ by matrix multiplication of the feature maps and attention maps: $\mathbf{K} = \mathbf{P}\mathbf{A}_p^T$. Here, $\mathbf{P}$ and $\mathbf{A}_p$ are temporarily reshaped to $\mathbb{R}^{c\times u}$ and $\mathbb{R}^{k^2\times u}$ ($u = s \times h \times w$) for conciseness.

Suppress Operation (SO). In this step, SBM leverages the above generated kernel to suppress the salient regions in $\{\mathbf{Q}_i\}_{i=s}^{t-1}$ which have been activated in $\mathbf{P}$. Then the later frames can pay attention to broader regions and obtain more complete representations.

In particular, a convolution operation is performed on the input feature map $\mathbf{Q}_i$ and the reshaped kernel $\mathbf{K}$ with size $c \times k \times k$:

$$\mathbf{R}_i = \mathrm{softmax}\left(\mathbf{Q}_i \star \mathbf{K}\right), \quad i = s, ..., t-1. \tag{3}$$

Here, $\star$ is convolution operation. $\mathrm{Softmax}$ operation is conducted on $h \times w$ dimension for normalization. The output $\{\mathbf{R}_i\}_{i=s}^{t-1} \in \mathbb{R}^{h\times w}$ is an affinity matrix, which presents high values for features captured in previous frames $\mathbf{P}$. Furthermore, to suppress the salient features with high similarities in $\mathbf{R}_i$, we inverse the affinity matrix. SO achieves this by a dedicated transformation, which transforms the affinity map $\mathbf{R}_i$ to the suppression matrix $\mathbf{S}_i$:

$$\mathbf{S}_i = \mathrm{e}^{\beta\left[\frac{1}{hw\mathbf{R}_i}-1\right]_-}, \quad i = s, ..., t-1. \tag{4}$$

Here, $[a]_- = \min\{a, 0\}$. The output matrix $\mathbf{S}_i \in \mathbb{R}^{h\times w}$ lies between 0 and 1. β is a hyperparameter to control the variance of the transformed distribution. A higher β represents heavier suppression, *i.e.*, giving lower weights to the salient features.

Then, SO multiplies $\mathbf{S}_i$ with $\mathbf{X}_i$ to generate the final suppressed feature maps: $\mathbf{Y}_i = \mathbf{S}_i \cdot \mathbf{X}_i$. In this process, the salient region will be suppressed by the multiplication with low weights in $\mathbf{S}_i$. Notably, convolution in Equation 3 actually measures the patch-wise affinities instead of point-wise ones. So the results $\mathbf{S}_i$ and $\mathbf{R}_i$ will be more continuous and smooth. Therefore, the multiplication will not incur the discontinuities in $\mathbf{Y}_i$, which may complicate the local relation modeling for later convolutions.

Overall, for the later frames, SBM decreases the saliency/attention of the salient region, and thus focuses on a broader region. Therefore, the attention regions of consecutive frames transit from salient to broader ones.

Cross Propagation Module (CPM). Salient features that are suppressed (or dropped) in SO are harmful for the later frames because they hinder the mining of less salient regions. However, for the entire video, the lost salient information is still ID-related and thus helpful. So, we use CPM to preserve them by transiting them to the unsuppressed frames $\{\mathbf{X}_i\}_{i=0}^{s-1}$ (*i.e.*, the red dashed lines in Figure 2). Hyperparameter α (0.1 by default) will control the transition's degree. The procedure is formulated as:

$$\mathbf{Z}_i = \mathbf{X}_i + \alpha \sum_{j=s}^{t-1} \mathrm{GAP}\left(\mathbf{X}_j - \mathbf{Y}_j\right), i = 0, ..., s-1. \tag{5}$$

Meanwhile, the information loss may also deteriorate the representations of frames $\{\mathbf{X}_i\}_{i=s}^{t-1}$ since the most salient features are suppressed. To increase representation capability, we also encourage the information flow from $\{\mathbf{X}_i\}_{i=0}^{s-1}$ to $\{\mathbf{Y}_i\}_{i=s}^{t-1}$ (the green dashed lines in Figure 2) as:

$$\mathbf{Z}_i = \mathbf{Y}_i + \alpha \sum_{j=0}^{s-1} \mathrm{GAP}(\mathbf{X}_j), \quad i = s, ..., t-1. \tag{6}$$

3.2. Integration and Distribution Module (IDM)

We also propose IDM to assist SBM. SBM is dedicated to enhancing the representational ability of video-level features. But the performance of SBM also depends heavily on the information richness of each frame's attention region. To be specific, the more informative these regions of every frame are, the more powerful the final video-level representations in SBM will be.

To enrich frame-level representations, recent methods [26, 41, 44] mainly adopt self-attention mechanisms, which enrich a position as a similarity-based aggregation with features of all positions. However, as pointed out in [38], the non-local behavior is much more crucial than the similarity-based aggregation strategy.

Inspired by this, IDM adopts an integration and distribution structure to approximate *input-agnostic* affinity maps and thus maintains the non-local behavior. In this way, IDM enables the message to pass across all frames and thus enriches the representations of frames.

Figure 3 gives an overview of our IDM. In detail, IDM will integrate key features/channels and then distribute them via a fixed mode. We firstly reshape the middle-level feature maps $\mathbf{X} \in \mathbb{R}^{c\times t\times h\times w}$ to $\mathbf{F} \in \mathbb{R}^{c\times m}$ ($m = t \times h \times w$). Then a series of linear transformations will be applied on $\mathbf{F}$ to enable the information flow between positions. Formally,

$$\mathbf{F}' = \mathbf{G}^{\mathbb{D}}\mathbf{G}^{\mathbb{I}}\mathbf{F}\left(\mathbf{L}^{\mathbb{I}}\mathbf{L}^{\mathbb{D}} + \mathbf{I}\right) \tag{7}$$

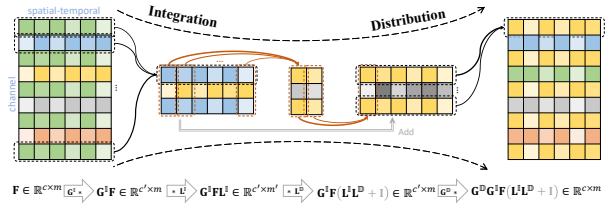

Figure 3. The framework of IDM. Matrix multiplication in thick arrow transforms the feature maps. The width of thin arrow indicates the aggregation weight.

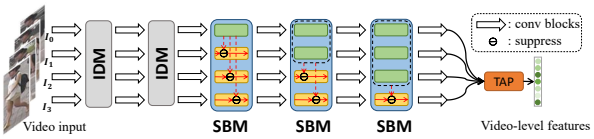

Figure 4. The architecture of our SINet. TAP denotes temporal average pooling. In SBM, the green and yellow blocks indicate the front and later frames. The later (yellow) frames will be suppressed with the guidance of the front (green) frames.

where $\mathbf{G}$ and $\mathbf{L}$ represent the linear transformation performed on the c (channel) and m (spatial-temporal) dimensions, respectively. Superscripts $\mathbb{I}$ and $\mathbb{D}$ denote the integration and distribution, respectively. In detail, $\mathbf{G}^{\mathbb{D}} \in \mathbb{R}^{c\times c'}$, $\mathbf{G}^{\mathbb{I}} \in \mathbb{R}^{c'\times c}$, $\mathbf{L}^{\mathbb{I}} \in \mathbb{R}^{m\times m'}$, $\mathbf{L}^{\mathbb{D}} \in \mathbb{R}^{m'\times m}$, and $m' \leq m$, $c' \leq c$ in general. Identity matrix $\mathbf{I} \in \mathbb{R}^{m\times m}$ denotes residual connection. The calculation order is shown in Figure 3, *i.e.*, the integration and distribution on m dimension are in the middle of the c dimension. The overall procedure of IDM is like an encoder-decoder, where integration operation extracts the key features, and distribution operation tries to recover the original feature maps with these key features. For easy combination with the backbone, the output of IDM has a residual connection [11]: $\mathbf{F}_{output} = \mathbf{F}' + \mathbf{F}$.

IDM can also be reviewed from the perspective of "affinity map". In fact, the matrix $\mathbf{L}^{\mathbb{I}}\mathbf{L}^{\mathbb{D}}$ in Equation 7 is a low-rank decomposition of $\mathbf{L} \in \mathbb{R}^{m\times m}$, an affinity map measuring the similarities of any two positions. Thus, $\mathbf{L}$ realizes the message passing between any two positions in the spatial-temporal dimension, so dose $\mathbf{G}$ for the channel dimension. In this way, IDM establishes the connections of arbitrary two positions in different frames or even different channels, and incurs powerful mutual enhancement between frames.

Our IDM shares several desirable advantages compared with previous self-attention methods [26, 38, 44]. First, the affinity map $\mathbf{L}$ (or $\mathbf{G}$) is learnable. Therefore, IDM can automatically discover useful patterns from the training distribution, such as the foreground and silhouette. Second, $\mathbf{L}$ (or $\mathbf{G}$) endows our IDM with the ability to subtract features (not only weighted addition) and increases the flexibility in composing features [35]. Third, IDM is more efficient as it optimizes the order of matrix chain multiplication compared with [38]. To be specific, the complexity reduces from $\mathcal{O}(m^2c')$ to $\mathcal{O}(mm'c')$.

IDM is reciprocal to SBM as they enhance the information richness of frames and diversities between frames, respectively. Notably, the salient-to-broad transition of SBM also enriches the information for later frames by broadening attention regions. However, the information gains are caused by the enlargement of the attention region, while the gains of IDM are incurred by the homogeneous message passing across frames without size variation.

Relations to GCNet. GCNet [2] with avg-pooling can be regarded as a special case of our IDM. Formally, GCNet replaces the $\left(\mathbf{L}^{\mathbb{I}}\mathbf{L}^{\mathbb{D}} + \mathbf{I}\right)$ in Equation 7 with all 1's matrix. Therefore, IDM has more complex modeling ability and better generality.

3.3. SINet for Video Person Re-ID

Network. Figure 4 shows the overview of our proposed SINet, which is a hierarchical combination of **S**BM and **I**DM. The backbone of SINet is ResNet-50 [11] pre-trained on ImageNet [6]. ResNet-50 has four layers and each layer is composed of several residual blocks. We insert two IDMs into the second layer after the middle and last residual blocks, respectively. Three SBMs are plugged evenly into the third layer and form a hierarchical structure to realize the salient-to-broad transition. In detail, the first SBM uses the first frame to suppress the remaining frames, while the second (third) SBM uses the first two (three) frames to extract salient features and performs suppression.

Finally, a temporal average pooling is used to generate the final video-level embeddings, which will be used in training or retrieval. We also adopt a batch normalization [20] to refine the embeddings as in [17].

Objective Function. Same as [47], we employ Cross-Entropy Loss $\mathcal{L}_{cent}$ and Batch Hard Triplet Loss [13] $\mathcal{L}_{tri}$ to jointly guide the training procedure. Moreover, to maintain the diversities between frames, inspired by the [15, 21], we leverage mutual information loss $\mathcal{L}_{mi}$ to minimize the mutual information of different frames' embeddings. More details about $\mathcal{L}_{mi}$ are in Supplementary Material.

The overall objective function of our SINet is a combination of the above three terms:

$$\mathcal{L}_{all} = \mathcal{L}_{cent} + \lambda_1\mathcal{L}_{tri} + \lambda_2\mathcal{L}_{mi} \tag{8}$$

where λ_1 and λ_2 are hyperparameters to control the influence of each loss function.

4. Experiments

4.1. Datasets and Evaluation Metrics

We perform comprehensive empirical studies on multiple video re-id datasets, *i.e.*, MARS [48], LS-VID [23],

Table 1. Performance (%) comparison with state-of-the-arts on MARS, LS-VID, iLiDS-VID, and PRID-2011 datasets. We separate these methods into several categories based on their utilizations of temporal relation. **TW** means Temporal Weighting, **3D** means 3D CNNs, **ME** means Mutual Enhancement, and **DA** means Difference Amplification.

Methods			MARS		LS-VID		iLiDS-VID	PRID-2011
			mAP	rank-1	mAP	rank-1	rank-1	rank-1
TW	SeeForest [49]	CVPR 17	50.7	70.6	-	-	55.2	79.4
	Snippet [3]	CVPR 18	76.1	86.3	-	-	85.4	93.0
	STA [9]	AAAI 19	80.8	86.3	-	-	-	-
3D	M3D [24]	AAAI 19	74.1	84.4	40.1	57.7	74.0	94.4
	AP3D [10]	ECCV 20	85.1	90.1	73.2	84.5	88.7	-
	STRF [1]	ICCV 21	86.1	90.3	-	-	89.3	-
ME	GLTR [23]	ICCV 19	78.5	87.0	44.3	63.1	86.0	95.5
	STGCN [42]	CVPR 20	83.7	90.0	-	-	-	-
	MGH [41]	CVPR 20	85.8	90.0	-	-	85.6	94.8
	MG-RAFA [44]	CVPR 20	85.9	88.8	-	-	88.6	95.9
	GRL [28]	CVPR 21	84.8	**91.0**	-	-	90.4	96.2
	DenseIL [12]	ICCV 21	**87.0**	90.8	-	-	92.0	-
Others	VRSTC [18]	CVPR 19	82.3	88.5	-	-	83.4	-
	AFA [4]	ECCV 20	82.9	90.2	-	-	88.5	-
	TCLNet [17]	ECCV 20	85.1	89.8	70.3	81.5	86.6	-
	BiCnet-TKS [16]	CVPR 21	86.0	90.2	75.1	84.6	-	-
	STMN [7]	ICCV 21	84.5	90.5	69.2	82.1	91.5	-
DA	**SINet(ours)**	-	86.2	**91.0**	**79.6**	**87.4**	**92.5**	**96.5**

iLiDS-VID [37], and PRID-2011 [14]. Similar to existing works [44, 45], we adopt the Cumulated Matching Characteristics (CMC) curve and mean Average Precision (mAP) as evaluation metrics.

4.2. Implementation Details

Sparse temporal sampling strategy [36] is used to generate a clip containing 4 frames. Frames are resized to 256×128. Every mini-batch has 32 clips corresponding to 8 identities (each identity has 4 clips). We use the Adam optimizer [22] with weight decay 0.0005. The initial learning rate is set to 0.0003 and decays by 0.1 at every 40 epochs. The training stage ends at the 160-th epoch. We also use random flipping and random erasing with a probability 0.5 for data augmentation. In SBM, kernel size k is set to 3, and β is set to 5 in default. λ_1 and λ_2 in Equation 8 are set to 1 and 0.01, respectively. In the test stage, we use all frames in units of 4-frame clips and obtain the final video feature by averaging all those clip-level representations. Cosine similarity is used for retrieval.

4.3. Comparison with State-of-the-art Methods

Table 1 shows comparison results of our method and state-of-the-art methods on four prevalent datasets. Results demonstrate the superiority of our method over existing methods. Furthermore, we make a detailed comparing analysis and draw several conclusions.

First, our method is obviously stronger than **TW**-based methods [9, 34]: 4.7% rank-1 and 5.4% mAP improvement on MARS. This proves the superiority of mutual enhancement and difference amplification over temporal weighting in the utilization of temporal relations. In other words, merely using temporal cues to re-weight each frame or region does not make full use of rich temporal information.

Second, our SINet also suppresses **3D**-based methods [1, 10, 24]. The superiority may be caused by the fact that both SBM and IDM in SINet can utilize the rich temporal relations between *any* two frames' feature maps. However, AP3D can only model local temporal relations and M3D performs enrichment in the final frame-level embeddings.

Third, SINet achieves superior or comparable performance with those methods based on **ME** [12, 23, 28, 41, 42, 44]. In fact, our SINet can enhance both video-level and frame-level representations with SBM and IDM, respectively. This complementarity incurs more powerful and informative video features. Our SINet performs slightly worse than DenseIL [12] on mAP. We attribute the inferiority to that DenseIL selects 8 frames for each video sequence in training, while we use only 4 frames due to the limitation of GPU memory.

Fourth, SINet also exceeds other methods, including recover-based [18], coherent-based [4] and erase-based [17]. Notably, TCLNet [17] can also be categorized in **DA**, as it aims to focus on the divergent attention regions in the later frames. However, TCLNet is inferior to our SINet since it erases the salient features in the later frames. This erasing operation deteriorates the representation capability of the final frames' embeddings. Conversely, our

Table 2. Performance (%) comparison for whether SBM or IDM is added to the baseline on MARS and LS-VID.

Methods	MARS		LS-VID	
	mAP	rank-1	mAP	rank-1
base.	83.2	88.6	73.3	82.8
+SBM	85.7	90.2	77.1	85.1
+IDM	85.9	90.5	78.0	86.2
SINet	**86.2**	**91.0**	**79.6**	**87.4**

SINet merely suppresses the salient features and maintains the completeness of representations.

In summary, SINet outperforms or is on a par with all those methods on four video re-id datasets, showing the generalization capability of our SINet on different scenes in terms of dataset scale, tracklet length, and resolution.

4.4. Ablation Study

For fair comparison, we build the baseline as the degraded SINet without SBM and IDM. We denote baseline as 'base.' for simplification.

Effectiveness of SBM and IDM. Table 2 illustrates the effectiveness of SBM and IDM. SBM brings 1.6% rank-1 and 2.5% mAP gains over the strong baseline on MARS. This shows the effectiveness of the salient-to-broad transition. IDM achieves comparable or even better improvement than SBM. The improvement proves that it is helpful to encourage message passing across arbitrary two positions. Our SINet, the combination of SBM and IDM, can further boost the rank-1 from 90.5% to 91.0% on MARS. The result validates the complementarity between SBM and IDM.

Component Analysis of SBM. Table 3 shows the effectiveness of the components in SBM. Specifically, both CAL and CPM have indispensible importance for the overall performance. This demonstrates the usefulness of channel selection in CAL and information propagation in CPM.

To evaluate the organization of SBMs, we insert three SBMs into the same position. As shown in Table 3, the result of 'Same Pos.' is clearly inferior to our SBM. We attribute the inferiority to that the hierarchical organization can leverage the semantic information at different levels.

We also evaluate the influence of kernel size k in KGM. As shown in Figure 5(a), in the beginning, the performance increased as the kernel size becomes larger, and reaches the peak when $k = 3$. We argue that a large kernel is more robust to filter the striking background. However, the performance decreases when $k > 3$. Actually, too large k may degrade performance because the corresponding kernel tends to smooth the response of salient and non-salient features, thus misguiding the selection of salient regions.

Component Analysis of IDM. Here we validate the influence of different numbers of IDMs. As shown in Table 3, adding more IDMs can improve performance from 89.8% to 90.5% in rank-1 accuracy. We believe that multiple IDMs can also conduct multi-hop communication as in [38].

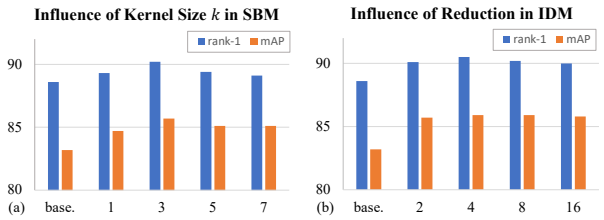

Figure 5. Results (%) on MARS illustrating the influences of different hyperparameters. (a) influence of kernel size k in SBM; (b) influence of reduction in IDM.

Table 3. Component analysis (%) of our SINet on MARS.

Methods	mAP	rank-1
base.	83.2	88.6
base. + Non-local [38]	85.1	89.7
base. + Non-local (**1**)	85.1	90.0
base. + Non-local (inverse)	85.0	89.7
base. + GC block [2]	85.0	89.8
base. + SBM w/o CAL	85.1	89.9
base. + SBM w/o CPM	85.4	90.0
base. + SBM (Same Pos.)	85.3	89.5
base. + SBM	85.7	90.2
base. + 1 IDM	85.7	89.8
base. + 2 IDM (Default)	85.8	90.5
base. + 3 IDM	85.8	90.5
SINet (base. + SBM + IDM)	**86.2**	**91.0**

The relation between c (m) and c' (m') is a vital factor to influence the performance of IDM. Figure 5(b) gives several comparisons, where numbers in x-axis denote the reduction c/c' and m/m'. We can observe that '4' achieves the best accuracy on both rank-1 and mAP. In fact, a large reduction will decrease the information richness after the integration operation, and a small reduction tends to overfit the training set as it has more parameters.

4.5. Comparison with Related Methods

In this section, we will compare our method with Non-Local(NL) block [38] and GC block [2]. We use the same configurations and compare them in three aspects:

Performance. Table 3 gives the performance comparisons of the above three methods. As we can see, SBM outperforms the NL block, validating the superiority of difference amplification over traditional mutual enhancement in utilizing temporal relations.

Table 3 also shows the performance of NL with several similarity-based aggregation strategies. Whether we inverse the affine matrix or fill the matrix with 1, the performances are similar, *i.e.*, the aggregation strategies are useless and dispensable. This further shows the superiority and ratio-

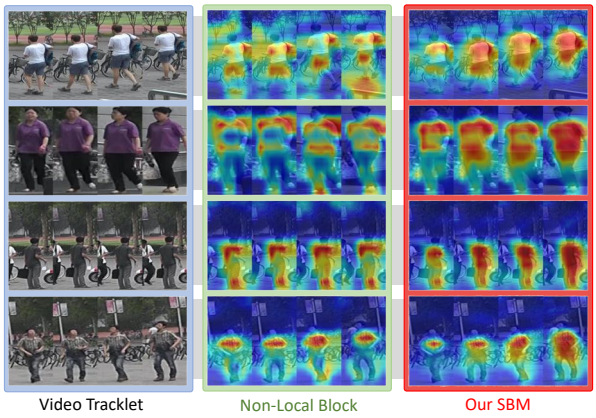

Figure 6. Attention regions of Non-Local and our SBM.

nality of our input-agnostic IDM.

GC block also achieves a favorable result without similarity calculation but is still inferior to our IDM. This verifies the effectiveness of IDM in using a more sophisticated and learnable matrix to model spatial-temporal relations.

Computational Complexity. Our IDM increases 1.081 GFlops computation cost over baseline, while NL incurs 2.216 GFlops. The cost is reduced by half due to the optimization of matrix chain multiplication. Notably, SBM increases negligible cost, *i.e.*, 0.014 GFlops, that mainly caused by the convolution and SE block.

Visualization. In Figure 6, we use CAM to visualize the attention regions of NL and SBM. Clearly, NL is prone to focus on some partial and identical regions across different frames, such as the upper clothes. Although these local features are usually salient, the overlook of other potential cues leads to limited and one-sided representations of pedestrians. Conversely, the attention regions of our SBM transit from salient towards broader ones. While the front frames have activated some salient regions as in NL, SBM will encourage the later frames to capture broader areas and mine other helpful cues. Overall, our SBM covers nearly the entire foreground, resulting in more complete and distinguishable embeddings. Notably, with our delicate SBM, the later frames always activate the foreground and will not introduce background noise. Meanwhile, the salient features are still salient as they are captured in all frames.

4.6. Visualization for Retrieval Results

To better understand the necessity of broad attention regions, Figure 7 shows some retrieval results on MARS. As can be observed, it is hard for the baseline to distinguish pedestrians that share similar salient regions, *e.g.*, the red shirt of the first person and the white shirt of the other two. In these cases, overdependence of salient regions may misguide the retrieval procedure.

Conversely, the salient-to-broad transition of SINet enlarges the attention regions. So, our SINet can successfully identify these pedestrians via other lesser salient but broader cues, *e.g.*, the existence of black socks, the difference between backpacks/lower clothes for three pedestrians, respectively. The incorporation of both salient and broad cues leads to diverse and integral characteristics of the given pedestrian and can rectify the false decisions made by only salient regions.

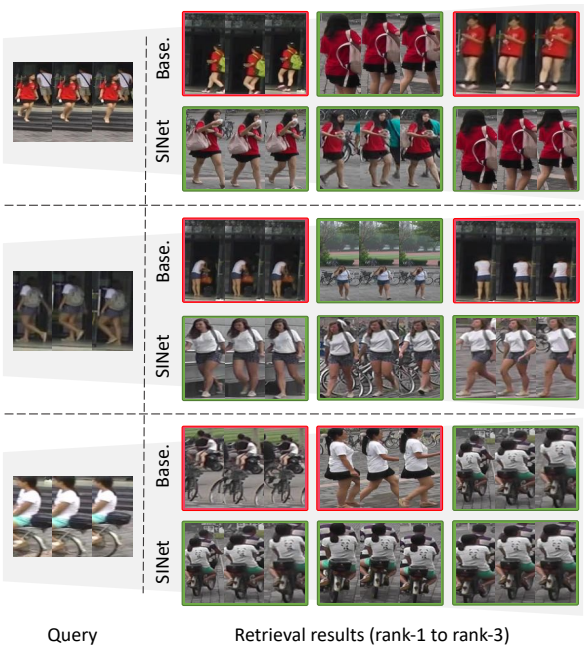

Figure 7. Visualization of retrieval results for the baseline and our SINet on MARS. The green and red backgrounds represent the correct and incorrect matches, respectively.

5. Conclusion

This paper aims to pursue a better representational capability for video person re-id. We present SBM to enlarge the attention regions for consecutive frames gradually. To further improve our SBM, we introduce IDM to consolidate frame-level representations. IDM and SBM are complementary and can be combined to form SINet. Extensive experiments show the effectiveness of our method.

Broader impacts. The proposed method boosts the performance of video re-id, making it more practicable in security, autonomous driving, and other security issues. Meanwhile, the higher accuracy and more used security cameras may raise the risk of privacy leaking and other security issues, which may put everyone under monitoring.

Acknowledgement This work is partially supported by Natural Science Foundation of China (NSFC): 61876171 and 61976203.

References

[1] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K. Roy-Chowdhury, and Ziyan Wu. Spatio-temporal representation factorization for video-based person re-identification. In *ICCV*, pages 152–162, 2021. 6

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *ICCV Workshops*, pages 0–0, 2019. 5, 7

[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018. 2, 6

[4] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, pages 660–676, 2020. 2, 6

[5] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, pages 3844–3852, 2016. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[7] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *ICCV*, pages 12036–12045, 2021. 6

[8] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *AAAI*, pages 3558–3565, 2019. 1

[9] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, volume 33, pages 8287–8294, 2019. 2, 6

[10] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, pages 228–243, 2020. 2, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 5

[12] Tianyu He, Xin Jin, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Dense interaction learning for video-based person re-identification. In *ICCV*, pages 1490–1501, 2021. 6

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[14] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011. 6

[15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 5

[16] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, pages 2014–2023, 2021. 6

[17] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, pages 388–405, 2020. 2, 5, 6

[18] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, pages 7183–7192, 2019. 2, 6

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 1, 3

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 5

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 5

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, pages 3958–3967, 2019. 2, 5, 6

[24] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, volume 33, pages 8618–8625, 2019. 2, 6

[25] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. 4

[26] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *BMVC*, 2019. 1, 2, 4, 5

[27] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *TCSVT*, 28(10):2788–2802, 2018. 2

[28] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, pages 13334–13343, 2021. 6

[29] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *CVPR*, pages 8766–8775, 2020. 2

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

[31] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *ICCV*, pages 6942–6950, 2019. 3

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 1

[33] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018. 1

[34] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, pages 562–572, 2019. 2, 6

[35] Yi Tay, Anh Tuan Luu, Aston Zhang, Shuohang Wang, and Siu Cheung Hui. Compositional de-attention networks. In *NeurIPS*, pages 6135–6145, 2019. 5

[36] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 6

[37] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014. 6

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1, 2, 4, 5, 7

[39] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, pages 5177–5186, 2018. 1

[40] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *ICCV*, pages 6589–6598, 2019. 2

[41] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *CVPR*, pages 2899–2908, 2020. 1, 2, 4, 6

[42] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, pages 3289–3299, 2020. 1, 2, 6

[43] Wei Zhang, Xiaodong Yu, and Xuanyu He. Learning bidirectional temporal cues for video-based person re-identification. *TCSVT*, 28(10):2768–2776, 2017. 2

[44] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, pages 10407–10416, 2020. 1, 4, 5, 6

[45] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 6

[46] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR*, pages 4913–4922, 2019. 2

[47] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *CVPR*, pages 8514–8522, 2019. 5

[48] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016. 5

[49] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 4747–4756, 2017. 2, 6