Event-guided Person Re-Identification via Sparse-Dense Complementary Learning

Chengzhi Cao[1], Xueyang Fu[1*], Hongjian Liu[1], Yukun Huang[1],
Kunyu Wang[1], Jiebo Luo[2], Zheng-Jun Zha[1]
[1]University of Science and Technology of China, China
[2]University of Rochester, USA
{chengzhicao@mail., xyfu@, jeffeey@mail., kevinh@mail., kunyuwang@mail.}ustc.edu.cn,
jluo@cs.rochester.edu, zhazj@ustc.edu.cn

Abstract

Video-based person re-identification (Re-ID) is a prominent computer vision topic due to its wide range of video surveillance applications. Most existing methods utilize spatial and temporal correlations in frame sequences to obtain discriminative person features. However, inevitable degradation, e.g., motion blur contained in frames, leading to the loss of identity-discriminating cues. Recently, a new bio-inspired sensor called event camera, which can asynchronously record intensity changes, brings new vitality to the Re-ID task. With the microsecond resolution and low latency, it can accurately capture the movements of pedestrians even in the degraded environments. In this work, we propose a Sparse-Dense Complementary Learning (SDCL) Framework, which effectively extracts identity features by fully exploiting the complementary information of dense frames and sparse events. Specifically, for frames, we build a CNN-based module to aggregate the dense features of pedestrian appearance step by step, while for event streams, we design a bio-inspired spiking neural network (SNN) backbone, which encodes event signals into sparse feature maps in a spiking form, to extract the dynamic motion cues of pedestrians. Finally, a cross feature alignment module is constructed to fuse motion information from events and appearance cues from frames to enhance identity representation learning. Experiments on several benchmarks show that by employing events and SNN into Re-ID, our method significantly outperforms competitive methods. The code is available at https://github.com/Chengzhi-Cao/SDCL.

*Corresponding author. This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62276243 and U19B2038, the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-025.

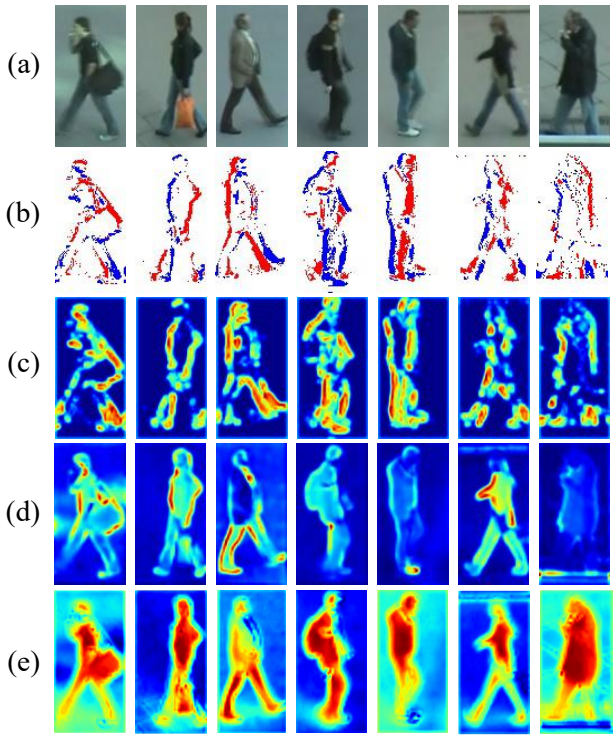

Figure 1. Visual examples of learned feature maps. From top to bottom: (a) original images, (b) corresponding events, (c) feature maps of events, (d) feature maps of frames in PSTA [49] (w/o events), (e) feature maps of frames in our network (w/ events).

1. Introduction

Person re-identification (Re-ID) identifies a specific person in non-overlapping camera networks and is used in a variety of surveillance applications [15, 31, 32]. Due to the availability of video data, video-based person Re-ID has attracted considerable attention. Compared with image-based Re-ID methods, video sequences contain numerous detailed

spatial and temporal information, which is beneficial to improving Re-ID performance [34, 46, 50].

Most existing video-based Re-ID approaches rely on spatial and temporal correlation modules, which are useful for deriving human representations that are resistant to temporal changes and noisy regions [19, 33, 51, 53]. To generate a person's representation from a video, they focus on shared information across numerous frames while taking into consideration the temporal context. Although video data can provide a wealth of appearance cues for identity representation learning, they also bring motion blur, illumination variations, and occlusions [24, 54]. These data-inherent phenomena result in the loss and ambiguity of essential identity-discriminating shape cues, and cannot be well solved by existing video-based Re-ID solutions [43].

Instead of depending solely on video sequences, this work intends to exploit event streams captured by event cameras to compensate for lost information and guide feature extraction in frames [44, 61]. Since the novel bio-inspired event camera can record per-pixel intensity changes asynchronously, it has high temporal resolution, high dynamic range, and low latency [12], providing a new perspective for person Re-ID. In other words, unlike traditional cameras that capture dense RGB pixels at a fixed rate, the event camera can accurately encode the time, location, and sign of the brightness changes [37, 41], offering robust motion information to identify a specific person.

In this paper, we propose a sparse-dense complementary learning network (SDCL) to fully extract complementary features of consecutive dense frames and sparse event streams for video-based person Re-ID. First, for dense video sequences, we build a CNNs-based backbone to aggregate frame-level features step-by-step. For sparse event streams, we design a deformable spiking neural network to suit the sparse and asynchronous characteristics of events. Because spiking neural network (SNN) has a specific event-triggered computation characteristic that can respond to the events in a nearly latency-free way, it is naturally fit for processing events and can preserve the spatial and temporal information of events by utilizing a discretized input representation. Meanwhile, we introduce deformable operation to deal with the degradation of spikes in deeper layers of SNN, better utilizing the spatial distribution of events to guide the deformation of the sampling grid. Finally, to jointly utilize sparse-dense complementary information, we propose a cross-feature alignment module to exploit the clear movement information from events and appearance cues from frames to enhance representation capacity. As shown in Figure 1, the feature maps of events still preserve the sparse distribution of events, which can guide the baseline to capture and learn discriminative representation clearly. Compared with the baseline (without events) in the fourth row, the learned feature maps in the fifth row show that our method tends to focus on the most important semantic regions in original frames and easily selects the better represented areas. The representation of events shows the contour and pose of a specific person. It presents that the sparse events can guide the baseline network to capture and learn discriminative representation clearly. the learned feature maps of dense RGB frames and sparse events intend to capture different semantic regions, but they still have spatial correlation. Both of them contribute to the final results.

This work makes the following contributions:

- We introduce a new modality, called event streams, and explore its dynamic properties to guide person Re-ID. To the best of our knowledge, this is the first event-guided solution to tackle the video-based Re-ID task.
- We propose a sparse-dense complementary learning network to fully utilize the sparse events and dense frames simultaneously to enhance identity representation learning in degraded conditions.
- We design a deformable spiking neural network to suit the sparse characteristics of event streams, which greatly utilizes the spatial consistency of events to provide motion information for dense RGB frames in a lightweight architecture.

Extensive experiments are conducted on multiple datasets to demonstrate how the bio-inspired event camera can help improve the Re-ID performance of baseline models and achieve higher retrieval accuracy than SOTA methods.

2. Related Work

Video-based Re-ID. Following the development of image-based Re-ID [6, 7, 23, 30, 38, 39, 48, 58], there are numerous recent progress in video-based Re-ID [14, 20, 40, 51, 55]. Most video-based Re-ID approaches aim to fully use spatial and temporal information and produce person representations that are resistant to a variety of factors, including human position, occlusion, and so on. To exploit the image-level features in a sequence-level person representation, Eom *et al.* [11] introduced a temporal memory module to save attentions that are optimized for typical temporal patterns in person videos. Wang *et al.* [49] explored the spatial correlations within a frame to determine the attention weight of each location and explored temporal consistency information to suppress the interference features and strengthen the discriminative ones. Aich *et al.* [1] presented a flexible new computational unit to extract complementary information along the spatio-temporal dimension.

Event-based Vision. Since events are bio-inspired visual signals that resemble the form of asynchronous spike trains, numerous bio-inspired learning methods are proposed for the event-based learning, such as recurrent neural

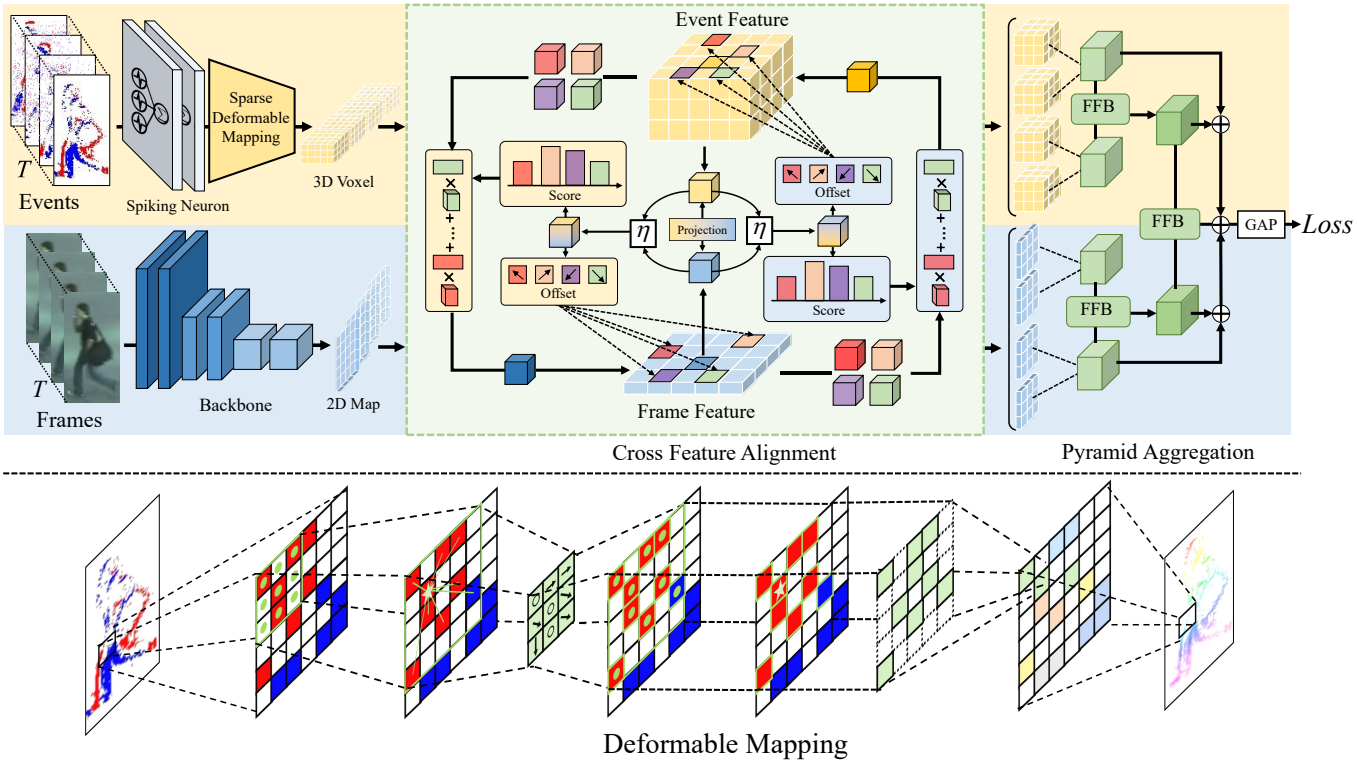

Figure 2. The overview of our Sparse-Dense Complementary Learning Network (SDCL), and we use four frames as an example. Given the frames, we first utilize ResNet-50 [17] as a backbone to extract frame-level features. For events, we deploy Spiking Neural Network (SNN) to preserve the spatial and temporal distribution of events and extract event-level features. With the guidance of events, the cross feature alignment module is utilized to compute the spatial consistency between frames and events, and then fuse them. Finally, a pyramid aggregation module is utilized to aggregate two types of features. "FFB" and operation "η" will be discussed later.

network [4, 9, 25]. There are also many CNNs-based methods to process event streams. Duan *et al*. [10] firstly deployed 3D U-Net and incorporated an E2I module to leverage HR image information to denoising and super-resolving events. They also implemented a display-camera system and proposed a multi-resolution event dataset. Paikin *et al*. [36] designed a three-phase architecture that fuses a conventional frame stream with the output of an event camera. Gehrig *et al*. [13] utilized cost volumes and introduces recurrency to incorporate temporal priors into dense optical flow estimation from event cameras. However, there still is no event-guided solution specifically designed for video-based person Re-ID.

3. Methods

3.1. Overview

As shown in Figure 2, our method takes the frame sequences with corresponding events in one timestamp T as the input. For RGB frames, they are fed into the ResNet backbone [17], whose parameters are trained for ImageNet Classification [28]) to extract features. For event streams, they are fed into SNN to extract spatial and temporal features of event streams simultaneously. Then, both frame features and sparse event features pass through a cross feature alignment module to extract complementary information contained in dense frames and sparse events. Finally, we employ a pyramid aggregation module to fuse two types of features ro enhance the identity representations. Below we detail each module of our SDCL.

3.2. Event Representation

When brightness change exceeds a predefined threshold c in a timestamp t, an event will be recorded as(x_t, y_t, p_t, t), where (x_t, y_t) and t denote the spatial and temporal location of the event, respectively. $p_t \in \{+1, -1\}$ represents the polarity of the brightness changes (increase or decrease). The polarity is computed by:

$$p_t = \Phi(log(\frac{\mathrm{L}_{xy}(t)}{\mathrm{L}_{xy}(t')}), c), \tag{1}$$

where c is the intensity threshold deciding whether an event can be recorded, $\mathrm{L}_{xy}(t)$ and $\mathrm{L}_{xy}(t')$ represent the instantaneous brightness intensity at time t and its previous time t', respectively. $\Phi(\cdot, \cdot)$ is a piece-wise function.

3.3. Deformable Spiking Neural Network

To suit the sparse characteristic of events, we deploy a spiking neural network to extract event features, and then propose deformable mapping operation to guide the deformation of convolutions and maintain the spatial information in events.

Spiking Model. The polarity of event streams represents an increase or decrease of brightness at one pixel. Inspired by the dynamics and adaptability of biological neurons, we choose the LIF model [42] to balance biological neurons' complicated dynamic features and their mathematical expressions. It is described as:

$$\tau_m \frac{dV(t)}{dt} = -V(t) + I(t), \tag{2}$$

where $V(t)$ and $I(t)$ denote the neuronal membrane potential and the pre-synaptic input at time t, τ_m is a constant. When $V(t)$ exceeds V_{th}, the neuron generates a spike and resets its membrane potential to its initial value. Here $I(t)$ is calculated as the weight sum of pre-spikes:

$$I(t) = \sum_{i=1}^{n} (w_i \sum_{k} \psi_i(t - t_k)), \tag{3}$$

where n denotes the number of pre-synaptic weights, w_i is the synaptic weight connecting i-th pre-neuron to post-neuron. $\psi_i(t - t_k)$ is a spike event from i-th pre-neuron at time t_k that k-th spike occurred. Its value is equal to 1 when $t = t_k$. The impact of each pre-spike is modulated by the corresponding synaptic weight (w_i) to generate a current influx to the post-neuron. Thus, the forward propagation is:

$$x_i^{t+1,n} = \sum_{j=1}^{n-1} w_{ij}^n o_j^{t+1,n-1}, \tag{4}$$

$$u_i^{t+1,n} = u_i^{t,n} f(o_i^{t,n}) + x_i^{t+1,n} + b_i^n, \tag{5}$$

$$o_i^{t+1,n} = g(u_i^{t+1,n}), \tag{6}$$

where n and t denote the nth layer and timestamp, w_{ij} is the learning weight from the jth neuron in pre-synaptic layer to the ith neuron in the post-synaptic layer. $f(x) = \tau e^{-\frac{x}{\tau}}$, and $g(\cdot)$ is a piece-wise function:

$$g(x) = \begin{cases} 1, & x \geq V_{th} \\ 0, & otherwise \end{cases} \tag{7}$$

Most of recent works [42] [12] have shown that the number of spikes drastically vanishes at deeper layers, resulting in serious performance degradation (as shown in Figure 4). It clearly limits the application of SNN in computer vision.

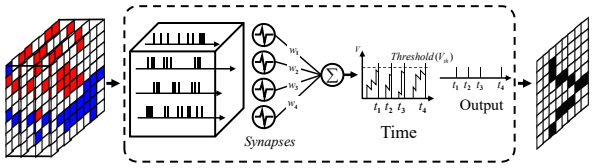

Figure 3. The structure of Leaky Integrate and Fire (LIF) Spike Neuron. The synaptic weight modulates the pre-spikes, which are then incorporated as a current influx in the membrane potential and decay exponentially. The post-neuron fires a post-spike and resets the membrane potential whenever the membrane potential reaches the firing threshold.

Event Deformable Mapping. To tackle abovementioned issue, previous work [29] has demonstrated that a hybrid SNN-CNN architecture can retain performance. But event stream is asynchronous and spatially sparse, so there are lots of margins in it. The traditional convolution, widely used in RGB images, is not suitable to extract event features. Motivated by the deformable convolutional network [8, 60], we utilize the sparse distribution of events to guide the deformation of convolutions to tackle degradations in SNN. Traditional deformable convolution is not suitable for event streams since the offsets are learned from the preceding feature maps and RoIs, but there are also some useless regions in event streams. So we intend to utilize the distribution of events to directly guide the deformation. As shown in Figure 2, we define the grid G as the receptive field. Take standard 2D convolution for example, they calculate the neighboring pixels but some of them don't have events, so most of calculation is useless. Firstly, We define a key events (x_0, y_0), and calculate the nearest events with the same number as standard kernel, and return the coordinates of them $\{(x_n, y_n)|n = 1, \dots N\}$. When we find the nearest events, our grid G will change its shape and totally cover them. This operation can be calculated as:

$$Y(m) = \sum_{n=1}^{N} w_n \cdot X(m + m_n + \triangle m_n), \tag{8}$$

where $m_n \in \{(-1. - 1), (-1, 0), ..., (1, 1)\}$ represents the locations in grid G, $\{\triangle m_n|n = 1, ..., N\}$ are the offsets, w is the learnable weight, and $N = |G|$. $Y(m)$ and $X(m)$ represent the features at location m from the input and output feature maps, respectively. We first choose a single event as the center, and employ manhattan distance to capture the neighboring events $\{x_n, y_n\}$:

$$L_n = |x_n - x_0| + |y_n - y_0|, \tag{9}$$

then we choose the nearest position of events as the sampling locations. We will enumerate all spatial locations in

the event voxel. The offset can be calculated as:

$$\triangle m_n(x_n) = x_n - x_0 - m_n(x_n), \tag{10}$$

$$\triangle m_n(y_n) = y_n - y_0 - m_n(y_n). \tag{11}$$

In the event voxel, the center of deformable convolution m_0 and the sampling regions $m_n + \triangle m_n$ will enumerate the location of all events. In this way, our deformable convolution only calculates the regions of events without consideration of margins in voxels.

3.4. Cross Feature Alignment

The aim of this module is to effectively aggregate dense frame features and sparse event features. The frames are fed into a backbone, such as ResNet-50 [17] to obtain 2D feature map, while the event sequences are processed by SNN to generate 3D voxel features. Although many existing attention operations can be directly used to aggregate these heterogeneous features, they still have limitations. For example, it is hard to keep the spatial consistency since the marginal regions in event sequences have great influence on the final feature maps. So we dynamically choose several key-point regions on the 3D voxel for each frame feature and vice versa To extract complementary features of frames and events. We design a symmetrical alignment structure so that each mode can learn extra attributes from the other one.

Assuming F_i is the image feature in i-th position, given a 2D feature map $\mathbf{F} = \{F_1, F_2, ..., F_{HW}\}$ and voxel features $\mathbf{P} = \{P_1, P_2, ..., P_J\}$, the reference points $R_i = (r_x^i, r_y^i)$ in the image plane are calculated from voxel feature $P_i = (p_x^i, p_y^i, p_z^i)$ as follows:

$$R_i = T_{e-f} \cdot P_i, \tag{12}$$

where T_{e-f} is the projection from events to frames. The query feature Q_i are computed as an element-wise product of the frame feature F_i and its corresponding event feature P_j. The final operation $\eta(\cdot, \cdot)$ can be calculated by:

$$\eta(Q_i, R_i) = \sum_{m=1}^{M} W_m [\sum_{i=1}^{I} MLP(Q_i) \cdot W'_m \mathbf{F}(R_i + \triangle R)], \tag{13}$$

where W_m and W'_m represent learnable weights, and MLP is a multilayer perceptron to generate attention scores. M and I are number of attention heads and sampling positions, respectively. $\triangle R$ is the sampling offset. The same operation in Eq. 13 is also employed to choose key-point regions on 2D feature maps for each voxel feature. This complementary operation can not only conduct cross-domain relational modeling with the help of dynamically generated sampling offset but also maintain position consistency between events and frames to obtain the reference points. By regarding it as a multi-modal learning method, each mode can actually learn some extra attributes from other modalities. After cross-feature alignment, we mutually enhance the feature of events and frames at lower stages of the network. Both event features and frame features contribute to the final results.

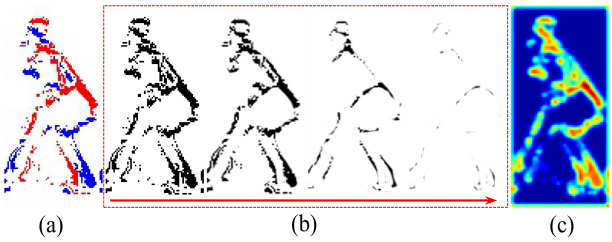

Figure 4. Visualization of features in SNN and deformable mapping. (a) presents events; (b) shows that the deeper the SNN layer is, the more the number of spikes vanishes. But using deformable mapping (c) can still preserve spatial information of events.

3.5. Pyramid Aggregation

To further exploit the spatial and temporal correlation from two types of modalities, and fuse information from both modes, we deploy a pyramid structure to aggregate two types of features. We first utilize a convolutional layer to transform the event voxel with the same size as frame features. Then, the adjacent frame features and event features will be fed into feature fusion block (FFB). We adopt STAM [49] module as FFB to obtain hierarchical features with temporal receptive fields. The final output features from this hierarchical architecture will pass through Global Average Pooling layer. The symmetrical multi-stage structure can not only maintain the spatial consistency between events and frames, but also obtain long-range temporal dependence from them.

3.6. Loss Function

Triplet loss and identification loss are widely used in person Re-ID. Following [1], we adopt the cross-entropy loss with label smoothing as the identification loss, and add batch triplet loss with hard mining strategy to optimize out network. The total loss $\mathcal{L}_{total}$ is the combination of two losses:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{tri}, \tag{14}$$

where λ_1 and λ_2 are the weights of two losses.

4. Experiments

In this section, we provide some implementation details and show ablation studies as well as visualization to evaluate the performance of our network. We compare our model with several state-of-the-art approaches, including OSNet [59], SRS-Net [45], STMN [11], STGCN [53], CTL [33], GRL [35], SINet [3], RAFA [56], MGH [52], TCLNet [21], STRF [2] and PSTA [49].

Table 1. Comparison on PRID, iLIDS-VID and MARS. Numbers in bold indicate the best performance and underscored ones are the second best. For the input, "V" and "E" represent that the input are the image sequences and event sequences, respectively. Results in brackets are obtained with the source codes provided by the authors.

Methods		PRID-2011		iLIDS-VID		MARS	
Network	Input	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
GRL [35]	V	92.7	89.9	90.1	84.7	82.2	88.3
OSNet [59]	V	92.7	89.9	89.0	82.7	81.4	87.3
SRS-Net [45]	V	88.8	84.3	89.8	84.0	82.9	88.1
STMN [11]	V	92.8	88.8	84.1	77.3	81.8	88.3
CTL [33]	V	91.5	87.6	84.2	77.3	82.7	89.3
PSTA [49]	V	92.3	88.8	88.1	80.0	83.1	89.2
STGCN [53]	V	-	-	-	-	83.7	90.0
SINet [3]	V	-	96.5	-	92.5	86.2	91.0
RAFA [56]	V	-	95.9	-	88.6	85.9	88.8
MGH [52]	V	-	94.8	-	85.6	85.8	90.0
TCLNet [21]	V	-	-	-	86.6	85.1	89.8
STRF [2]	V	-	-	-	89.3	86.1	90.3
GRL [35]	E	21.4	11.2	30.2	18.0	27.7	16.7
OSNet [59]	E	22.2	10.1	27.9	16.7	30.9	19.3
SRS-Net [45]	E	17.2	9.0	32.7	19.3	20.9	10.0
STMN [11]	E	20.2	11.2	23.5	12.7	22.4	10.0
CTL [33]	E	20.4	13.5	28.4	18.0	25.6	12.7
PSTA [49]	E	22.2	12.4	22.4	10.0	22.7	12.0
GRL [35]	V+E	93.2	87.6	90.6	85.3	82.8	88.7
OSNet [59]	V+E	93.7	89.9	90.1	84.7	81.9	87.7
SRS-Net [45]	V+E	91.5	87.6	<u>90.7</u>	<u>86.7</u>	83.8	89.3
STMN [11]	V+E	94.0	91.0	87.2	81.3	83.4	89.0
CTL [33]	V+E	93.9	91.0	88.4	82.0	<u>85.3</u>	89.6
PSTA [49]	V+E	<u>94.7</u>	<u>93.3</u>	88.6	83.3	85.1	<u>89.9</u>
Ours	V+E	**96.9**	**96.5**	**93.2**	**92.7**	**86.5**	**91.1**

4.1. Datasets

Since there is no available event person Re-ID dataset, we generate events from three classical video-based datasets for Re-ID, including PRID-2011 [18], iLIDS-VID [47] and MARS [57]. Following [10, 16, 44], we apply a display-camera system and simulator V2E [22] to generate corresponding event sequences. More details can be found in the supplementary material.

4.2. Implementation Details

Our network is implemented based on Pytorch on an Intel i4790 CPU and one NVIDIA RTX 2080Ti GPU. The initialization of our CNN encoder is ImageNet-pretrained standard ResNet-50 [17]. For each video clip, we use a constrained random sampling approach to randomly sample frames from evenly divided 8 chunks. We train our network for a total of 500 epochs, starting with a learning rate of 0.0003 and decaying it by 10 every 200 epochs. Adam [27] optimizer is applied to update the parameters. During testing, the cosine similarity is used to measure the distance between the gallery and the query.

4.3. Comparison with other methods

Table 1 shows comparison results of our method and other SOTA methods on MARS [57], PRID-2011 [18] and iLIDS-VID [47]. Note that to demonstrate the superiority of complementary learning strategy over traditional video-based methods, we classify our experiments into three groups depending on the input data, including videos only, events only, and video with events. When we regard events as CNN's input, we deploy them into event voxels to encode the positive and negative events. These results illustrate that: **(1)** Since event streams can only capture the brightness change in the scene, they lost color-based information, such as content and saturation. Thus, using only events cannot achieve promising performance, with the best value of 30.9% mAP on MARS. **(2)** When the inputs are videos and events (V+E), we keep the video-based network structure unchanged and use a SNN to extract features of events, and then employ the spatial fusion block [5] into the mainstream. Compared with video-based methods, when events are utilized to guide video-based methods, their performance is mostly improved. For example, STMN (V+E) achieves an improvement up to 3.1% and 1.6% mAP on iLIDS-VID and MARS, respectively; CTL (V+E) achieves up to 3.4%, 4.7%, and 0.3% in terms of Rank-1 accuracy on PRID-2011, iLIDS-VID and MARS, respectively. The superiority is caused by the fact that events bring useful information to video-based methods. **(3)** Our method outperforms PSTA [49], with an improvement up to 2.2% and 1.4% mAP on PRID-2011 and MARS, respectively. The comparison clearly demonstrates the effectiveness of complementary fusion for exploring complementary information between events and frames.

Moreover, to demonstrate the robustness of our method in degraded conditions, we follow [26] to create blur and occlusion in PRID-2011 and iLIDS-VID dataset, and present the experimental results in Table 3. The first group is designed for dealing with RGB frames, while the second group is fed into frames and events simultaneously. Compared with the result in Table 1, we notice that blurry artifacts have great influence on the performance of all methods. For example, the mAP value of PSTA in PRID-2011 drops from 92.3% to 79.7%. But when event streams are fed into networks, it makes great progress in mAP by 5.3% on PRID-2011 and 12.5% on iLIDS-VID, respectively. In occluded condition, We can observe that it improves the 2nd best method SRS-Net [45] by 6.2% mAP accuracy in iLIDS-VID dataset. The comparison clearly demonstrates

Table 2. Quantitative results on different fusion methods. Note that "E2F" means utilizing event feature to guide frame feature extraction, while "F2E" means utilizing frame feature to guide event feature extraction.

Methods			PRID-2011		iLIDS-VID	
Fusion	E2F	F2E	mAP	Rank-1	mAP	Rank-1
-	-	-	85.3	78.7	85.2	80.0
Concat	-	-	84.6	77.5	84.0	77.3
Addition	-	-	67.8	57.3	56.2	46.1
Attention	✓		89.4	84.3	88.4	82.7
		✓	87.7	79.8	87.8	81.3
	✓	✓	90.6	85.4	90.6	85.3
Ours	✓		92.7	89.9	88.6	83.3
		✓	89.9	85.4	86.5	80.7
	✓	✓	**96.9**	**96.5**	**93.2**	**92.7**

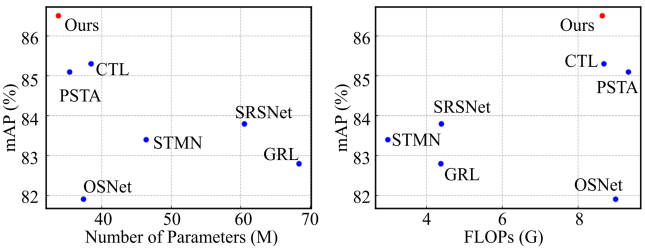

Figure 5. Parameters and FLOPs comparisons.

the superiority of events for exploring spatial and temporal correlations from videos.

4.4. Parameters and FLOPs

We list the parameters and FLOPs of all compared methods in Figure 5. It is clear that our SDCL has comparable storage consumption with consideration of acceptable FLOPs to achieve the highest accuracy. Note that SNN can improve energy efficiency and reduce paramters because it is only active when it receives or emits spikes, while CNNs operate with all units active regardless of real-valued input or output values. Since in the cross feature alignment module, we calculate the response as a weighted sum of the features at all positions between event features and frame features, it greatly increases FLOPs.

4.5. Ablation Study

Components of SNN. We evaluate the contribution of each component (including spiking neurons and deformable mapping) in PRID-2011 dataset to demonstrate the efficiency of our network. The results are shown in Figure 6. To demonstrate the degradation of SNN at deeper layers, we generally increase the number of SNN layers. It is obvious that when we set the number of SNN from 1 to 2, the mAP and Rank-1 accuracy increases, but when the number of SNN is 3, the performance drops slightly.

Table 3. The mAP values of different methods on PRID and iLIDS-VID dataset in degraded (blurry and occluded) conditions.

Methods		PRID-2011		iLIDS-VID	
Method	input	Blurry	Occluded	Blurry	Occluded
SRS-Net [45]	V	78.8	77.7	61.6	71.6
STMN [11]	V	81.5	72.9	57.7	62.8
GRL [35]	V	81.8	76.1	60.7	60.2
CTL [33]	V	81.6	73.8	56.0	70.4
PSTA [49]	V	79.7	71.3	58.9	67.5
SRS-Net [45]	V+E	82.1	82.7	67.8	74.4
STMN [11]	V+E	85.5	82.2	62.3	69.0
GRL [35]	V+E	88.4	83.0	66.2	73.5
CTL [33]	V+E	88.6	78.5	64.2	75.9
PSTA [49]	V+E	85.0	78.9	65.5	73.0
Ours	V+E	**89.5**	**88.9**	**71.4**	**80.7**

In addition, we also fix two SNN layers, and increase the number of deformable mapping in the following layers. The mAP and Rank-1 accuracy in PRID-2011 and iLIDS-VID datasets still increases continuously, which illustrates that our deformable mapping can maintain the sparse distribution in event streams to provide extra brightness information. However, keep increasing the number eventually brings limited improvement. Therefore, we select the number of deformable mapping as 3 in our experiments. More analysis can be found in the supplementary material.

Components of Cross Feature Alignment Module. To illustrate the effect of our cross feature alignment module, we compare it with several widely used operations, including concatenation, addition, attention mechanism, and our module. The results are shown in Table 2. It is clear that concatenation and addition have bad influence on PRID-2011 and iLIDS-VID datasets. This is because both of them directly combine sparse events and dense RGB frames without considering the difference between them. When using the attention mechanism, it achieves 90.6% mAP and 85.4% Rank-1 accuracy in iLIDS-VID, surpassing the baseline by a large margin. Finally, our cross feature alignment module outperforms the attention mechanism by 6.3% and 2.4% mAP in PRID-2011 and iLIDS-VID datasets, respectively. To illustrate the mutual information between events and frames, we also conduct experiments by only deploying our module in "E2V" (event-to-video) or "V2E" (video-to-event) to verify the complementary effect between events and frames, and experimental results are reported in Table 2. We can see that single complementary learning between events and frames can also increase the Rank-1 accuracy and mAP. With "E2V" complementary information, our network achieves 92.7% mAP and 89.9% Rank-1 accuracy in PRID-2011, higher by 2.8% mAP and 4.5% Rank-1 than "V2E", which means that the sparse distribution of events is more beneficial for RGB frames to get better per-

Table 4. Ablation study of pyramid aggregation. PA means pyramid aggregation, and Res denotes residual block adopted in PA.

Dataset		PRID-2011		iLIDS-VID	
Metric		mAP	Rank-1	mAP	Rank-1
w/o PA	-	89.2	83.1	88.1	82.6
w/ PA	Res	92.3	88.8	90.2	86.0
	FFB	**96.9**	**96.5**	**93.2**	**92.7**

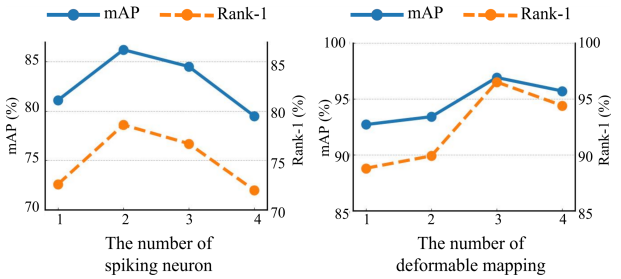

Figure 6. Quantitative ablation study on different components in Deformable Spiking Neural Network.

formance. It can be explained that dense RGB frames contain abundant global contextual information and physical connections of human body, but event streams can also provide extra brightness information for video-based Re-ID.

Pyramid Aggregation. We evaluate the pyramid aggregation by aggregating features with residual block to demonstrate the benefits of feature fusion block (FFB). The Table 4 shows that the pyramid aggregation structure can actually obtain better performance.

4.6. Visualization

In Figure 1, we visualize the learned feature maps in our network. It shows that the event features is able to preserve the sparse distribution of events and provide accurate pose and shape information to locate specific person. Since the feature maps of dense RGB frames and sparse events focus on specific semantic regions, both contribute to the final performance improvement.

Moreover, we visualize the learned feature maps extracted by the baseline and our methods under blurry and occluded conditions in Figure 7. For occluded frames, the features of baseline are more related to the boundaries of occlusion masks, while our method still focus on the representative areas of each person. This is because the event camera does not record static occlusion without changes in position or brightness, causing the event data to guide the baseline network to focus more on moving objects. For blurry frames, since event streams contain intensity variations with high temporal resolutions, motion blur effect can be precisely represented. As shown in Figure 7 (c), the

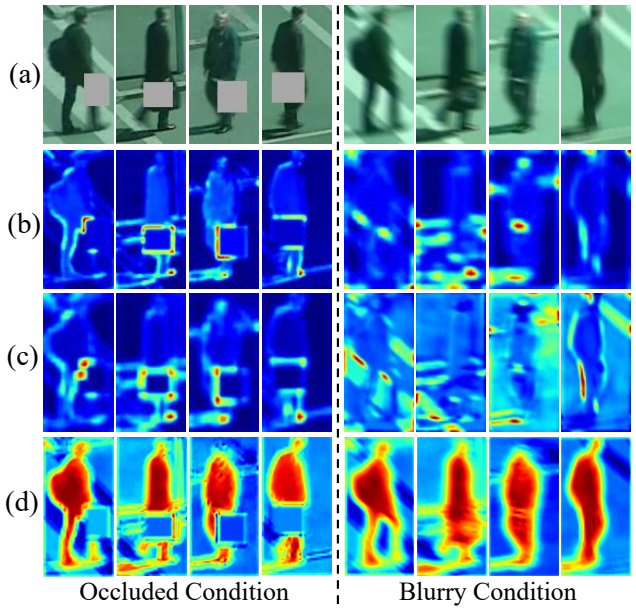

Figure 7. Visual examples of the learned feature maps in blurry and occluded conditions. (a) shows original occluded images and blurry images, (b) is the feature maps of frames in PSTA [49], (c) and (d) represent feature maps of frames in our network (w/o and w/ events, respectively).

feature maps of baseline (without events) pay more attention to the sidewalk boundaries, which cannot extract correct features of pedestrians easily. On the contrary, our method (with events) can learn more discriminative information from events, making the network more robust to different perturbations.

5. Conclusion

In this paper, we explore the benefit of events to guide video-based Re-ID. We proposed a novel Sparse-Dense Complementary Learning (SDCL) that learns complementary spatio-temporal representations from frames and events to deal with video-based Re-ID tasks. One advantage of our method over most video-based Re-ID models is that we firstly deploy sparse event streams into dense frame-level features to fully utilize the extra brightness information to guide video-based baselines. To help baseline models discover more discriminative spatio-temporal representations for robust video Re-ID, we design a deformable SNN that extracts event-level features while preserving spatially sparse distribution of events. Moreover, we propose a complementary learning module to capture comprehensive clues contained in dense frames and sparse events. Extensive experiments on benchmarks have shown the benefit of events for video-based Re-ID.

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, 2021. 2, 5

[2] 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. 5, 6

[3] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In CVPR, pages 7339–7348, 2022. 5, 6

[4] Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous Optical Flow and Intensity Estimation from an Event Camera. In CVPR, 2016. 3

[5] Chengzhi Cao, Xueyang Fu, Yurui Zhu, Gege Shi, and Zheng-Jun Zha. Event-driven video deblurring via spatio-temporal relation-aware network. 6

[6] Xiaodong Chen, Xinchen Liu, Wu Liu, Xiao-Ping Zhang, Yongdong Zhang, and Tao Mei. Explainable Person Re-Identification with Attribute-guided Metric Distillation. In ICCV, 2021. 2

[7] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta Batch-Instance Normalization for Generalizable Person Re-Identification. In CVPR, 2021. 2

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable Convolutional Networks. 4

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In ICCV, 2015. 3

[10] Peiqi Duan, Zihao W. Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. EventZoom: Learning to Denoise and Super Resolve Neuromorphic Events. In CVPR, 2021. 3, 6

[11] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based Person Re-identification with Spatial and Temporal Memory Networks. In ICCV, 2021. 2, 5, 6, 7

[12] Guillermo Gallego, Tobi Delbrück, and Garrick Orchard. Event-Based Vision: A Survey. 2021. 2, 4

[13] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras. 2021. 3

[14] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-Preserving 3D Convolution for Video-Based Person Re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, ECCV. 2020. 2

[15] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal Knowledge Propagation for Image-to-Video Person Re-Identification. In ICCV, 2019. 1

[16] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. EvIntSR-Net: Event Guided Multiple Latent Frames Reconstruction and Super-Resolution. 2021. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 3, 5, 6

[18] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person Re-identification by Descriptive and Discriminative Classification. In Anders Heyden and Fredrik Kahl, editors, Image Analysis. 2011. 6

[19] 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, 2021. 2

[20] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal Complementary Learning for Video Person Re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, ECCV. 2020. 2

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

[22] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2e: From Video Frames to Realistic DVS Events. 6

[23] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing Status Awareness for Long-Term Person Re-Identification. In ICCV, 2021. 2

[24] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In CVPR, pages 14084–14094, 2020. 2

[25] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In CVPR, 2017. 3

[26] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In CVPR, pages 18963–18974, 2022. 6

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

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017. 3

[29] Chankyu Lee, Adarsh Kumar Kosta, Alex Zihao Zhu, Kenneth Chaney, Kostas Daniilidis, and Kaushik Roy. Spike-FlowNet: Event-Based Optical Flow Estimation with Energy-Efficient Hybrid Neural Networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, ECCV. 2020. 4

[30] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined Depth Space based Architecture Search For Person Re-identification. In CVPR, 2021. 2

[31] Jianing Li, Shiliang Zhang, Jingdong Wang, Wen Gao, and Qi Tian. Global-Local Temporal Representations for Video Person Re-Identification. In ICCV, 2019. 1

[32] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-identification. In CVPR, 2018. 1

[33] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. Spatial-Temporal Correlation and Topology Learning for Person Re-Identification in Videos. In CVPR, 2021. 2, 5, 6, 7

[34] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang. A Spatio-Temporal Appearance Representation for Video-Based Pedestrian Re-Identification. In ICCV, 2015. 2

[35] 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, 2021. 5, 6, 7

[36] Genady Paikin, Yotam Ater, Roy Shaul, and Evgeny Soloveichik. EFI-Net: Video Frame Interpolation from Fusion of Events and Frames. In CVPRW, 2021. 3

[37] Liyuan Pan, Cedric Scheerlinck, and Yuchao Dai. Bringing a Blurry Frame Alive at High Frame-Rate With an Event Camera. In CVPR, 2019. 2

[38] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by Aligning: Visible-Infrared Person Re-identification using Cross-Modal Correspondences. In ICCV, 2021. 2

[39] Min Ren, Lingxiao He, Xingyu Liao, Wu Liu, Yunlong Wang, and Tieniu Tan. Learning Instance-level Spatial-Temporal Patterns for Person Re-identification. In ICCV, 2021. 2

[40] Minho Shim, Hsuan-I Ho, Jinhyung Kim, and Dongyoon Wee. READ: Reciprocal Attention Discriminator for Image-to-Video Re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, ECCV. 2020. 2

[41] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the Sim-to-Real Gap for Event Cameras. 2020-08-22. 2

[42] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. Neural networks, 111:47–63, 2019. 4

[43] Rahul Rama Varior, Gang Wang, Jiwen Lu, and Ting Liu. Learning invariant color features for person reidentification. IEEE Transactions on Image Processing, 25(7):3395–3410, 2016. 2

[44] Bishan Wang, Jingwei He, and Wen Yang. Event Enhanced High-Quality Image Recovery. 2020-07-16. 2, 6

[45] Haoran Wang, Licheng Jiao, Shuyuan Yang, Lingling Li, and Zexin Wang. Simple and effective: Spatial rescaling for person reidentification. IEEE Transactions on neural networks and learning systems, 2020. 5, 6, 7

[46] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person Re-identification by Video Ranking. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, ECCV. 2014. 2

[47] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person Re-identification by Video Ranking. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, ECCV. 2014. 6

[48] Xueping Wang, Shasha Li, Min Liu, Yaonan Wang, and Amit K. Roy-Chowdhury. Multi-Expert Adversarial Attack Detection in Person Re-identification Using Context Inconsistency. In ICCV, 2021. 2

[49] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid Spatial-Temporal Aggregation for Video-based Person Re-Identification. In ICCV, 2021. 1, 2, 5, 6, 7, 8

[50] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly Attentive Spatial-Temporal Pooling Networks for Video-Based Person Re-identification. In ICCV, 2017. 2

[51] 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, 2020. 2

[52] 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. 5, 6

[53] 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, 2020. 2, 5, 6

[54] Yi-Fan Zhang, Hanlin Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. Learning domain invariant representations for generalizable person re-identification. arXiv preprint arXiv:2103.15890, 2021. 2

[55] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-Based Person Re-Identification. In CVPR, 2020. 2

[56] 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. 5, 6

[57] 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 Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, ECCV, 2016. 6

[58] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online Pseudo Label Generation by Hierarchical Cluster Dynamics for Adaptive Person Re-identification. In ICCV, 2021. 2

[59] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In ICCV, pages 3702–3712, 2019. 5, 6

[60] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable ConvNets V2: More Deformable, Better Results. 4

[61] Dongqing Zou. Learning Event-Driven Video Deblurring and Interpolation. In Andrea Vedaldi and Horst Bischof, editors, ECCV. 2020. 2