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

Chengzhi Cao¹, Xueyang Fu¹, Hongjian Liu¹, Yukun Huang¹, Kunyu Wang¹, Jiebo Luo², Zheng-Jun Zha¹
¹University of Science and Technology of China, China
²University 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
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 recurrence 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 to 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:

$$p_t = \Phi(\log(\frac{L_{xy}(t)}{L_{xy}(t')}), c),$$

where $c$ is the intensity threshold deciding whether an event can be recorded. $L_{xy}(t)$ and $L_{xy}(t')$ 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:

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

where $V(t)$ and $I(t)$ denote the neuronal membrane potential and the pre-synaptic input at time $t$, $\tau_m$ 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 sum of pre-spikes:

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

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:

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

$$u_i^{t+1,n} = u_i^{t,n} f(a_i^{t,n}) + x_i^{t+1,n} + b_i^n,$$

$$a_i^{t+1,n} = g(u_i^{t+1,n}),$$

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

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

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.

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 $\{(x_i, y_i) | n = 1, \ldots, N\}$. When we find the nearest events, our grid $G$ will change its shape and totally cover them. This operation can be calculated as:

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

where $m_n \in \{(-1, -1), (-1, 0), \ldots, (1, 1)\}$ represents the locations in grid $G$, $\{\triangle m_n | n = 1, \ldots, N\}$ 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 $\{x_n, y_n\}$:

$$L_n = |x_n - x_0| + |y_n - y_0|,$$

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:
\[
\begin{align*}
\triangle m_n(x_n) &= x_n - x_0 - m_n(x_n), \\
\triangle m_n(y_n) &= y_n - y_0 - m_n(y_n).
\end{align*}
\]

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 \(F = \{F_1, F_2, ..., F_{HW}\}\) and voxel features \(P = \{P_1, P_2, ..., P_J\}\), 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:
\[
R_i = T_{e-f} \cdot P_i,
\]

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 \(P_i\). The final operation \(\eta(\cdot, \cdot)\) can be calculated by:
\[
\eta(Q_i, R_i) = \sum_{m=1}^{M} W_m \sum_{i=1}^{I} MLP(Q_i) \cdot W_{m}^r F(R_i + \triangle R),
\]

where \(W_m\) and \(W_{m}^r\) 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.

### 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 \(L_{total}\) is the combination of two losses:
\[
L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{tri},
\]

where \(\lambda_1\) and \(\lambda_2\) 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>PRID-2011</th>
<th>iLIDS-VID</th>
<th>MARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>Input</td>
<td>mAP Rank-1</td>
<td>mAP Rank-1</td>
</tr>
<tr>
<td>GRL [35]</td>
<td>V</td>
<td>92.7 89.9</td>
<td>90.1 84.7</td>
</tr>
<tr>
<td>OSNet [59]</td>
<td>V</td>
<td>92.7 89.9</td>
<td>89.0 82.7</td>
</tr>
<tr>
<td>SRS-Net [45]</td>
<td>V</td>
<td>88.8 84.3</td>
<td>89.8 84.0</td>
</tr>
<tr>
<td>STMN [11]</td>
<td>V</td>
<td>92.8 88.8</td>
<td>84.1 77.3</td>
</tr>
<tr>
<td>CTL [33]</td>
<td>V</td>
<td>91.5 87.6</td>
<td>84.2 77.3</td>
</tr>
<tr>
<td>PSTA [49]</td>
<td>V</td>
<td>92.3 88.8</td>
<td>88.1 80.0</td>
</tr>
<tr>
<td>STGCN [53]</td>
<td>V</td>
<td>93.7 89.9</td>
<td>89.6 84.0</td>
</tr>
<tr>
<td>SINet [3]</td>
<td>V</td>
<td>95.6 92.5</td>
<td>92.5 92.6</td>
</tr>
<tr>
<td>RAFA [56]</td>
<td>V</td>
<td>95.9 88.6</td>
<td>85.9 88.8</td>
</tr>
<tr>
<td>MGH [52]</td>
<td>V</td>
<td>94.8 85.6</td>
<td>85.8 90.0</td>
</tr>
<tr>
<td>TCLNet [21]</td>
<td>V</td>
<td>93.3 86.6</td>
<td>85.1 89.8</td>
</tr>
<tr>
<td>STRF [2]</td>
<td>V</td>
<td>89.3 89.3</td>
<td>86.1 90.3</td>
</tr>
<tr>
<td>GRL [35]</td>
<td>E</td>
<td>21.4 11.2</td>
<td>30.2 18.0</td>
</tr>
<tr>
<td>OSNet [59]</td>
<td>E</td>
<td>22.2 10.1</td>
<td>27.9 16.7</td>
</tr>
<tr>
<td>SRS-Net [45]</td>
<td>E</td>
<td>17.2 9.0</td>
<td>32.7 19.3</td>
</tr>
<tr>
<td>STMN [11]</td>
<td>E</td>
<td>20.2 11.2</td>
<td>23.5 12.7</td>
</tr>
<tr>
<td>CTL [33]</td>
<td>E</td>
<td>20.4 13.5</td>
<td>28.4 18.0</td>
</tr>
<tr>
<td>PSTA [49]</td>
<td>E</td>
<td>22.2 12.4</td>
<td>22.4 10.0</td>
</tr>
<tr>
<td>GRL [35]</td>
<td>V+E</td>
<td>93.2 87.6</td>
<td>90.6 85.3</td>
</tr>
<tr>
<td>OSNet [59]</td>
<td>V+E</td>
<td>93.7 89.9</td>
<td>90.1 84.7</td>
</tr>
<tr>
<td>SRS-Net [45]</td>
<td>V+E</td>
<td>91.5 87.6</td>
<td>90.7 86.7</td>
</tr>
<tr>
<td>STMN [11]</td>
<td>V+E</td>
<td>94.0 91.0</td>
<td>87.2 81.3</td>
</tr>
<tr>
<td>CTL [33]</td>
<td>V+E</td>
<td>93.9 91.0</td>
<td>88.4 82.0</td>
</tr>
<tr>
<td>PSTA [49]</td>
<td>V+E</td>
<td>94.7 93.3</td>
<td>88.6 83.3</td>
</tr>
<tr>
<td>Ours</td>
<td>V+E</td>
<td><strong>96.9</strong></td>
<td><strong>96.5</strong></td>
</tr>
</tbody>
</table>

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 i7-9700 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>PRID-2011</th>
<th>iLIDS-VID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion</td>
<td>E2F</td>
<td>F2E</td>
</tr>
<tr>
<td>Concat</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Addition</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Attention</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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 parameters 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>PRID-2011</th>
<th>iLIDS-VID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method input</td>
<td>Blurry</td>
<td>Occluded</td>
</tr>
<tr>
<td>SRS-Net [45] V</td>
<td>78.8</td>
<td>77.7</td>
</tr>
<tr>
<td>STMN [11] V</td>
<td>81.5</td>
<td>72.9</td>
</tr>
<tr>
<td>GRL [35] V</td>
<td>81.8</td>
<td>76.1</td>
</tr>
<tr>
<td>CTL [33] V</td>
<td>81.6</td>
<td>73.8</td>
</tr>
<tr>
<td>PSTA [49] V</td>
<td>79.7</td>
<td>71.3</td>
</tr>
<tr>
<td>SRS-Net [45] V+E</td>
<td>82.1</td>
<td>82.7</td>
</tr>
<tr>
<td>STMN [11] V+E</td>
<td>85.5</td>
<td>82.2</td>
</tr>
<tr>
<td>GRL [35] V+E</td>
<td>88.4</td>
<td>83.0</td>
</tr>
<tr>
<td>CTL [33] V+E</td>
<td>88.6</td>
<td>78.5</td>
</tr>
<tr>
<td>PSTA [49] V+E</td>
<td>85.0</td>
<td>78.9</td>
</tr>
<tr>
<td>Ours V+E</td>
<td>89.5</td>
<td>88.9</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PRID-2011</th>
<th>iLIDS-VID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>mAP Rank-1</td>
<td>mAP Rank-1</td>
</tr>
<tr>
<td>w/o PA</td>
<td>- 89.2</td>
<td>88.1</td>
</tr>
<tr>
<td>w/ PA</td>
<td>Res 92.3</td>
<td>90.2</td>
</tr>
<tr>
<td>FFB</td>
<td>96.9</td>
<td>93.2</td>
</tr>
</tbody>
</table>

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

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).

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 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


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


[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


[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


[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


[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


[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


[56] 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

[57] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kajian 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

[58] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Max Welling. editors, ECCV. 2, 5, 6

