Event-driven Re-Id: A New Benchmark and Method Towards Privacy-Preserving Person Re-Identification

Shafiq Ahmad[1,2] Gianluca Scarpellini[1,3] Pietro Morerio[2] Alessio Del Bue[2,3]
{shafiq.ahmad, gianluca.scarpellini, pietro.morerio, alessio.delbue}@iit.it
[1]Universita degli Studi di Genova, Italy, [2]Pattern Analysis & Computer Vision (PAVIS) - Istituto Italino di Tecnologia,
[3]Visual Geometry and Modelling (VGM) - Istituto Italiano di Tecnologia

Abstract

The large-scale use of surveillance cameras in public spaces raised severe concerns about an individual privacy breach. Introducing privacy and security in video surveillance systems, primarily in person re-identification (re-id), is quite challenging. Event cameras are novel sensors, which only respond to brightness changes in the scene. This characteristic makes event-based vision sensors viable for privacy-preserving in video surveillance. Integrating privacy into the person re-id; this work investigates the possibility of performing person re-id with the event-camera network for the first time. We transform the asynchronous events stream generated by an event camera into synchronous image-like representations to leverage deep learning models and then evaluate how complex the re-id problem is with this new sensor modality. Interestingly, such event-based representations contain meaningful spatial details which are very similar to standard edges and contours. We use two different representations, image-like representation and their transformation to polar coordinates (which carry more distinct edge patterns). Finally, we train a person re-id model on such images to demonstrate the feasibility of performing event-driven re-id. We evaluate the performance of our approach and produce baseline results on two synthetic datasets (generated from publicly available datasets, SAIVT and DukeMTMC-reid).

1. Introduction

Person re-identification (re-id) aims at recognizing the same person across multiple non-overlapping camera views. Re-id has gained significant interest in the computer vision community as being an enabling technology for intelligent video surveillance systems (e.g., tracking in non-overlapping views, forensic and security applications [41, 17]). The person re-id problem has been extensively studied in standard (RGB) camera networks and the advent of deep-learning-based re-id approaches [36, 41] has improved the performance rapidly. Most of such existing re-id models are developed for conventional RGB cameras, although some methods have been proposed for cross-modal re-id between RGB and infrared [7] or depth frames [2].

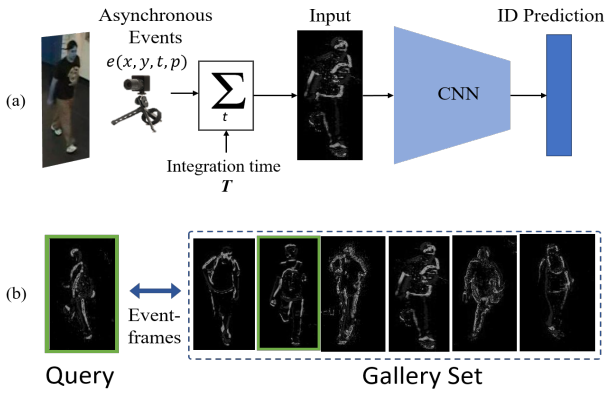


Figure 1. The proposed event-based person re-id system: (a) shows the block diagram where the output of event-camera converts to event-frame and then fed to the deep learning-based re-id network to perform person re-identification; (b) presents an example of query image (event-frame) matching with gallery set (in green).

Because of the growing demand for intelligent video surveillance systems, security and privacy concerns are coming closer together as parallel requirements. However, unauthorized access to the video surveillance data captured with traditional vision sensors is a severe threat to individuals' privacy. It is essential to make surveillance data secure from misuse, where it may be used for identity theft, blackmail, and mass surveillance [8]. According to the European General Data Protection Regulation (GDPR), article 5(1)(b) states: "Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes."

In recent years, event cameras have attracted attention of the computer vision community due to their working

principle and advantages such as high frame rates, high dynamic range (HDR), and no motion blur. Unlike standard cameras, which capture frames at a fixed frame-rate, event cameras record asynchronous brightness changes of a scene (called events), which substantially decreases the sensor's latency and extends its applications in surveillance. Besides naturally discarding redundant visual information, event cameras are indeed a feasible option for such visual surveillance applications where privacy-preserving is required. As without image appearance, it can better guarantee the anonymity of the subjects that can lead to solving the privacy concerns of vision applications in public spaces. To address the privacy-related issues, in current literature various approaches are suggested based on standard vision sensors, e.g., masking the human subjects [39] or blurred the detected faces [8] in the video, selective video surveillance method [11] and other image encryption techniques [12]. One of the main drawbacks is that it is difficult and complicated to ensure end-to-end privacy with these hand-designed techniques. Nevertheless, an event camera is an adequate substitute to prevent invasion of the individual's privacy in visual surveillance. In addition, as opposed to standard cameras, event-based vision sensors are quite efficient to work in varying or low illumination conditions and record fast movement.

Given that event cameras output stream of sparse and asynchronous event data, traditional image-based re-id algorithms cannot be applied directly. A methodological replacement is to develop probabilistic filters and spiking neural networks (SNN) to process event data [1, 25]. Alternatively, groups of events are converted into intermediate input representations with a regular - synchronous - tensor-like structure which are compatible with conventional vision algorithms [14, 21, 34]. This work is the first attempt to address the opportunities and the challenges of the privacy preserving re-id problem in event-based vision. Initially, we construct two synthetic dataset from publicly available benchmarks through event simulators [13, 29]. Next, utilizing group of events (2D histogram or event-frame) to perform person matching in non-overlapping event-camera views as shown in Fig 1. We can observe that event-frame representation contains meaningful information similar to edges while the subjects are moving, as event cameras only record intensity changes and thus mainly encoding the boundaries of the person shape.

Hence, the redundant visual information is discarded, and the proposed re-id model only relies on motion-triggered events. Our experiments validate that person re-id is possible with such nominal spatial information. Moreover, recent advancements have shown that it is indeed possible to recover grayscale images from event streams using event-to-image reconstruction techniques [3, 24, 28, 30, 37, 6]. These works certainly are considered as computational attacks to the privacy-preserving properties of event sensors. Additionally, we implement person re-id on recovered grayscale images (via E2VID [30]) to inspect the impact of these computational attacks on the privacy preserving capabilities of event cameras.

The contributions of this works are as follows:

- We investigate if person re-identification using event cameras is viable. To the best of our knowledge, this work is the first attempt at deploying an event-based solution for such vision task.
- We propose a re-id pipeline which accumulates events into frames that are processed by a Polar Transformation and then fed to a Convolutional Neural Network (CNN).
- Since re-id datasets with event cameras do not exists, we propose two synthetic event-based person re-id datasets to show the validity of our methodology. These are generated from RGB datasets: DukeMTMC-reid [32] and SAIVT [5] by means of an open-source event simulator [13, 29], which have proven extremely effective in other vision tasks in past works.
- We explore the influence of computational attacks on privacy-preserving attributes of event sensors through multiple experiments; trained person re-id model on reconstructed grayscale images from event-stream and compare their results with those obtained on standard RGB/grayscale images.

The rest of the paper is organized as follows: section 2 reviews related literature in the event-based vision and person re-id; section 3 details the proposed methodology; experimental details and results are provided in section 4, while conclusions are drawn in section 5.

2. Related Work

Person Re-Identification: In the last decade, researchers have worked extensively on person re-id with conventional cameras, and re-id models based on deep learning have shown significant progress [36, 41, 43]. Nevertheless, current approaches still have a strong image texture bias. Shape/contour sketch-based re-id methods [26, 40] have been proposed to overcome dependency on color information. Pang et al. [26] introduced sketch re-id with a focus on cross-domain feature learning. Their approach matches professional artists' sketches with their RGB counterparts. Yang et al. [40] demonstrated the use of person contour sketch images for re-id but under moderate clothing change. Our work might have similarities with contour sketch [40], but it presents substantial differences. First, Yang et al. only focus on cross-cloth person re-id. Second, they employ an edge detector to generate contour sketch images. On the other hand, our approach employs event-cameras to capture moving targets directly and efficiently.

Privacy-Preserving: The widespread usage of visual surveillance in public places putting people's privacy at risk [8]. Strict data privacy regulations such as the California Consumer Privacy Act (CCPA) and the GDPR exist to prevent the possible misuse of data. Though, illegal access to visual data is a serious threat to a privacy breach. The research community developed various techniques to secure surveillance systems and visual data from recognizing individuals and unauthorized access [11, 12, 8, 27]. Alem et al. [11] implement selective surveillance that contained only aggressive and suspicious behavioral patterns video frames by introducing a dynamic chaotic image enciphering scheme, which enables frame encryption.

Julia et al. [8] and Marina et al. [27] introduce privacy-preserving in person re-identification. Juila at al. create anonymized datasets using face blurring on publicly available person re-id benchmark and demonstrate that data can be safely anonymized by blurring faces without compromising the performance of person re-id. In comparison, Marina et al. proposed person re-id system that uses an RGB-D camera in a top-view configuration to extract anthropometric features for the recognition of people to address both occlusions and privacy-preserving problems. The technique in [8] is hand-designed that is not able to ensure end-to-end privacy and the RGB-D camera top-view arrangement in [27] is not feasible in a practical environment. On the other hand, event cameras are considered inherently privacy-preserving due to their working principle.

One could posit that the event-based sensor modality is a step towards privacy-preserving vision but recent advancements have shown that it is indeed possible to recover grayscale images from event streams using patch-based dictionaries [3], variational models [24], photometric constancy [28], and deep learning-based solutions [30, 37, 6]. These works indeed are considered computational attacks to the privacy-preserving properties of the event camera. Note that this security aspect has already been detected and addressed in a recent work [9] that provides a method with a dedicated encryption framework for event-based stream.

Event-based Vision: Event cameras are a relatively recent vision modality useful for several computer vision applications, from low-level vision (e.g., objects detection and tracking [15, 22], and optical flow [4, 42]) to high-level vision (e.g., image reconstruction [31], segmentation [35], and recognition [1, 25]). The research community developed two main approaches to process events-based data streams: *(i)* methods for event-by-event, which use events as an asynchronous stream, such as probabilistic filters and spiking neural network (SNN) [1, 25]; *(ii)* methods that group events into image-like tensors that are then processed by image-based learning methods (DNNs, SVMs, Random Forests) [14, 21, 34]. Recent literature has promoted the latter (named group-of-events) as the most effective method for event-based vision tasks. Such advantage is given by the image-like representation that carries spatial information about scene edges, which are the most informative regions in standard (RGB) images. Moreover, it allows the employment of existing deep learning techniques, and it achieves remarkable results on several vision benchmarks using traditional machine learning methods [14, 21].

In [14], authors proposed a framework to convert event streams into grid-based representations for end-to-end learning, which yields an improvement on the object recognition task. Maqueda et al. [21] processed event-frames to predict steering angle for self-driving cars. Lagorce et al. [19] proposed classification model based on image-like representation called time-surface to classify 36 characters(0-9, A-Z). Scarpellini et al. [33] developed a pipeline for 3D human pose estimation in event-based vision by accumulating events stream into frame data and applied CNN to predict body keypoints. In addition, Wang et al. [38] developed human gait recognition approach in event-based vision. Nevertheless, the person re-identification problem has never been studied in event-based vision yet.

In this work, we try to answer whether person re-identification in event camera networks is feasible. Since event cameras naturally respond to moving edges in the scene, synchronous event-frames contain meaningful spatial details (e.g., edges and contours) while lacking redundant visual information. Building upon the insights of above mentioned methods, we exploit event-frame (contained edges and contour information) to tackle the person re-identification problem.

3. Methodology

Our approach aims at solving person re-id with event-cameras. Person re-id is a classical computer vision problem, but the research community lacks dataset captured with event-cameras to tackle this task. Moreover, conventional machine learning techniques (e.g. CNN) developed for image-based data can not be applied directly on asynchronous events. Section 3.1 describe how to synthesize events data from classical person re-id datasets. Section 3.2 explains a standard methodology for converting asynchronous events into event frames to enable learning with traditional CNN architectures.

3.1. Synthetic Event Generation

Event Generation Model: An event camera trigger event $\mathbf{e}_k = (x_k, y_k, t_k, p_k)$ independently whenever magnitude of the log brightness at pixel $\mathbf{v}_k = (x_k, y_k)^T$ and time t_k exceeds a predefined threshold, $C > 0$ as:

$$\Delta L(\mathbf{v}_k, t_k) = L(\mathbf{v}_k, t_k) - L(\mathbf{v}_k, t_k - \Delta t_k) \geq p_k C, \quad (1)$$

where the polarity of the event p_k is a Boolean value (± 1) and Δt is the time since the last event at $\mathbf{v}_k$. In a given

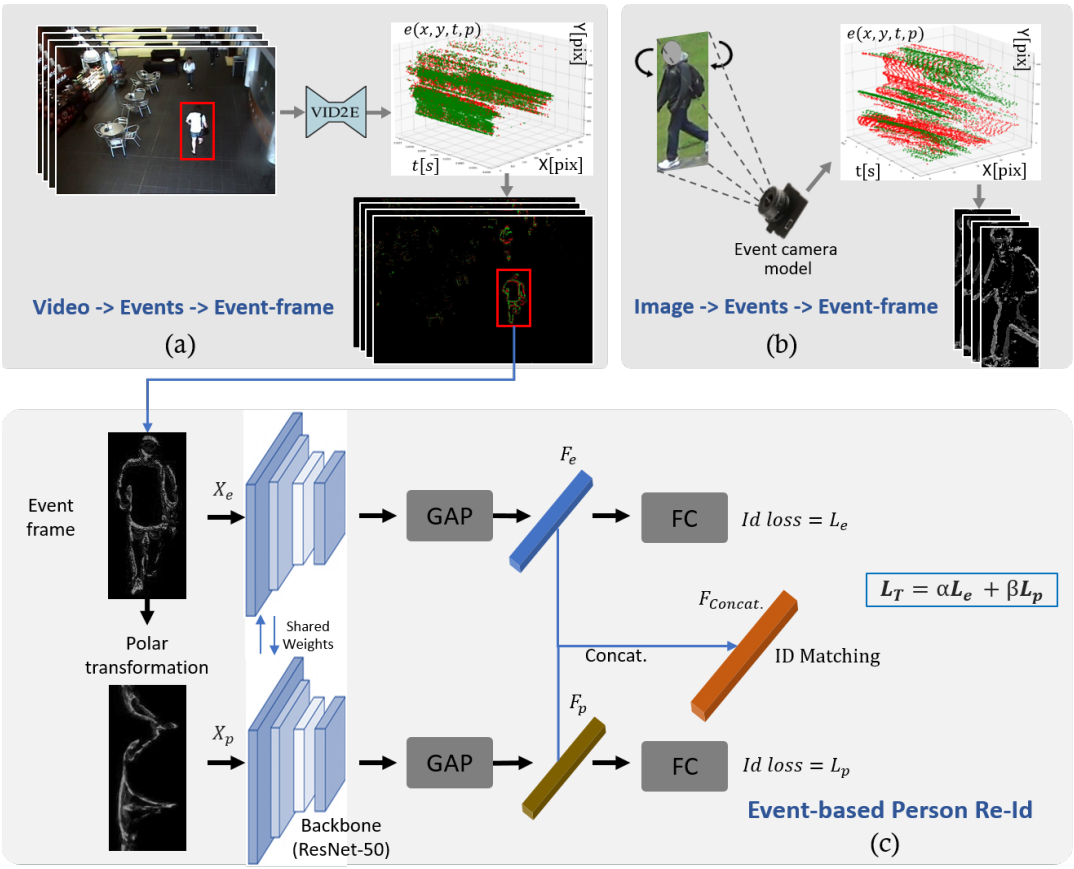


Figure 2. The complete pipeline of the proposed method, the existing re-id dataset is converted to synthetic events using an event simulator. (a) SAIVT video dataset is transformed using Vid2e simulator (b) DukeMTMC-reid image data using ESIM by applying an event-camera model on homographic movements of the image plane. Then in (a) & (b) asynchronous events are accumulated polarity pixel-wise into 2D histogram over constant time interval **T**. (c) ResNet-50 backbone with global average pooling (GAP) and a classifier as re-id network that is jointly trained on the event and polar images. At test time, the classifier is stripped off and a concatenated feature vector F_{Concat} is utilized for ID prediction

time interval δt , the event camera will activate a number of events:

$$E = \{e_k\}_{k=1}^{N} = \{(x_k, y_k, t_k, p_k)\}_{k=1}^{N}. \quad (2)$$

Event Generation: Due to the novelty of event sensor, event-based datasets are hardly available. To boost new research using event-cameras for different tasks, various event simulators [13, 29, 23] can transform RGB videos and images in most datasets into events streams. In this study, we applied simulators that act upon single image and videos. For classical image-based datasets (e.g DukeMTMC-reid), we adopt the open-source event simulator ESIM [29]. This method simulates homographic movements of the image plane and applies an event-camera model to synthesize simulated events. For video-based datasets (e.g. SAIVT), we adopt the event-simulator Vid2e [13]. Vid2e upsamples standard videos using a CNN and generates events from upsampled videos. We set the predefined threshold, C for both positive and negative events to $C_p = C_n = 0.1$.

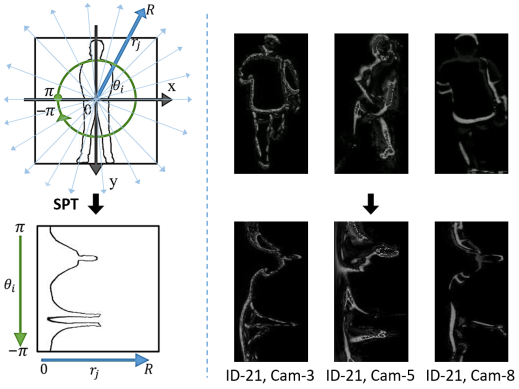
3.2. Event Representation and Learning

Event-to-Frame Conversion: Because of their asynchronous nature, events are represented as a set of points in a four-dimensional manifold comprised of spatial coordinates (x, y), time, and polarity. To process events with a CNN, it is necessary to convert the asynchronous events into a grid-like representation. Hence, we convert the asynchronous and sparse event stream into synchronous frames called "Event-frame" in order to leverage image based CNN architecture. Then, we accumulate the events polarity pixel-wise into a 2D histogram $H\pm$ [21] of both negative and positive events, using $p_k = \pm 1$, over a constant time interval given by:

$$H_{\pm}(x, y) = \sum_{t_k \epsilon T, p_k = \pm 1} \delta(x - x_k, y - y_k), \quad (3)$$

where δ is the Kronecker delta, and T is the time interval. Note that [21] preserves polarity while it converts events

into two-channel event images. On the contrary, we discard temporal as well as polarity information and produce single channel event image. The output image carries spatial information of scene edges, as shown in Fig. 2a. In the next step, the person re-id network is trained on generated synthetic event-frames and original bounding box labels from the RGB dataset. To achieve this, we choose the time interval (T≈40ms) to accumulate events which leads up to the time stamped ground truth label following [13] and then train the re-id model.



(a) (b)

Figure 3. Illustration of Spatial Polar Transformation (SPT). (a) example of SPT with uniform sampling angle θ_i; (b) top row, event-image of same identity in three different camera-view, below their transformation to polar coordinates.

Spatial Polar Representation: Since event cameras have a strong response to motion in the scene, we assume that person edges information in event-frame can be related to the person contour sketch. Inspired by Yang et al. [40], we transform an event-frame into polar coordinates. Thus, we aim at learning discriminant features from edge patterns as shown in Fig. 3. The size of the transformed image is set equal to the event image. Unlike the method in [40], we only use uniform sampling for transformation to polar coordinates to examine its influence on event-driven re-id. Another option would be to learn specific polar transformation, but [40] already demonstrated negligible improvements.

3.3. Person Re-Id

To unlock the potential of CNN architectures for our problem, we pre-process the output of the event camera (synthetic events). Initially, we partitioned the event stream into batches of events, and then each batch aggregated to build event-frames. Afterward, we employ ResNet-50 [18] as the backbone image feature embedding. The network takes two input images, event-frame X_e and its transformation to polar coordinates X_p of size 384 × 128. Thus, our model is trained jointly on events and polar images. The feature map of the last residual block feeds into the global average pooling layer. Finally, to predict the identity of the input person, we input the extracted feature vectors (F_e, F_p) into a classifier, which consists of a fully-connected (FC) layer and a Softmax loss function such as:

$$F_e = GAP(Conv_\theta(X_e)) \tag{4}$$

$$F_p = GAP(Conv_\theta(X_p)). \tag{5}$$

Further, we apply two classification losses (event loss and polar loss) to facilitate the learning feature F_e from event-frame and feature F_p from edge patterns in polar image respectively. Therefore, the overall loss of our model is,

$$L_T = \alpha L_e + \beta L_p, \tag{6}$$

where L_e and L_p represent the cross-entropy loss of ID classification of feature F_e and F_p respectively, both α and β are coefficients which control the contribution of the each loss. We empirically set these coefficients to $\alpha = 1$ and $\beta = 0.3$.

At test time, the classifier is stripped off and to perform person re-id our network encodes the query image by feature $F_{Concat.}$ as:

$$F_{Concat} = Concat(F_e, F_p), \tag{7}$$

obtained through concatenation of learned features F_e and F_p, which is applied for matching the gallery ones via nearest neighbour search (in Euclidean distances).

4. Experimental Setup and Results

4.1. Datasets

As no event-based re-id datasets are available, we generate synthetics events from two classical image-based datasets for re-id, SAIVT [5] and DukeMTMC-reid [32]. SAIVT is our main benchmark, having annotated sequences of 152 IDs, total 64,472 images captured from eight surveillance cameras. The dataset is recorded in an uncontrolled environment, each identity may only appear in a subset of cameras and images are subject to changes on viewpoints, illumination, and background. Most of the state-of-the-art methods evaluate this dataset for two camera pairs: cameras 3/8 (SAIVT-38) and cameras 5/8 (SAIVT-58), as the majority of the persons appeared in these two pairs. SAIVT-38 consists of 99 IDs, whereas SAIVT-58 from a dissimilar view, consists of 103 IDs. Furthermore, we also test on DukeMTMC-reid that contains 16,522 training images from 702 IDs, 17,661 gallery images from another 702 IDs, and 2,228 query images from the same IDs as query set. In addition to the viewpoint variations. DukeMTMC-reid main challenges are occlusions and people at lower resolutions.

Table 1. Rank accuracy and mAP (%) of the proposed method (using different representation) on SAIVT dataset.

SAIVT-38	R1	R5	R10	mAP
Ours$_{Event}$	63.6	77.9	85.1	50.2
Ours$_{Polar}$	56.5	73.5	81.6	39.0
Ours$_{Event+Polar}$	**73.5**	**84.7**	**89.8**	**55.3**
SAIVT-58				
Ours$_{Event}$	21.6	33.3	46.1	14.5
Ours$_{Polar}$	18.6	26.5	35.3	09.9
Ours$_{Event+Polar}$	**24.6**	**37.8**	**47.1**	**15.5**

Table 2. Performance comparison of our proposed method with baseline (RGB) on SAIVT dataset in terms of rank accuracy and mAP(%).

SAIVT-38	R1	R5	R10	mAP
Baseline$_{RGB}$	87.8	93.9	97.0	82.8
Ours$_{Event+Polar}$	73.5	84.7	89.8	55.3
SAIVT-58				
Baseline$_{RGB}$	53.9	65.7	72.6	51.2
Ours$_{Event+Polar}$	24.6	37.8	47.1	15.5

4.2. Experimental Setup

Evaluation protocol: After the conversion of SAIVT dataset, We obtained 8,110 event-frames for SAIVT-38 and 6,642 for in SAIVT-58. In both pairs, we randomly split the IDs into training and testing; in SAIVT-38, we take event-frames corresponding to 50 IDs for the training set and 49 IDs for the test set while, in SAIVT-58 52 IDs for the training set and 51 IDs for the test set. In the test set, we pick one query image for each Id in each camera and put the remaining images in the gallery. Besides, for DukeMTMC-reid, we acquired five event-frames for each image in the dataset. In the training set, we use all five event-frames while randomly selecting one event-frame for each image in the original gallery and query set. For evaluating our proposed method, we outline rank-1, rank-5 and rank-10 accuracy, and mAP for SAIVT dataset.

Implementation details: We use ResNet-50 as our backbone. Following [16], we apply GAP on the feature map from the last residual block and a linear layer (FC+BN+ReLU) to compute a 256-D feature embedding. We use ResNet trained with softmax loss and set the stride from 2 to 1 of the last stage. All training images are resized to 384 × 128 and then augmented by horizontal flip and normalization [16]. We set the batch size to 32 and train the model with a base learning rate 0.05 for 30 epochs and 0.005 for 60 epochs. We set momentum $\mu = 0.9$ and the weight decay to 5×10^{-4}. Our implementations are based on PyTorch framework with a single NVIDIA GeForce GTX 1180 GPU.

Table 3. Performance comparison of our proposed method with baseline on DukeMTMC-reid dataset in terms of rank accuracy.

Method	R1	R5	R10
Baseline$_{RGB}$	79.9	89.8	92.2
Ours$_{Event}$	16.6	26.1	31.8
Ours$_{Polar}$	06.6	12.0	17.1
Ours$_{Event+Polar}$	17.0	26.4	31.8

4.3. Results and Discussions

We conduct several experiments on the synthetic event dataset: SAIVT and DukeMTMC-reid and compare results with baseline.

SAIVT: Table 1 shows the experimental results on SAIVT-38 and SAIVT-58 (subsets of SAIVT) datasets for different settings. The model is trained jointly on two representations event-frame and its transformation to polar coordinates. Additionally, the model is trained on each single representation to validate the effectiveness individually. From Table 1 the following observations can be drawn:
(i) **Performance of event-frame vs polar image.** The impact utilizing single representation suggest that model trained with event-frame consistently perform better than polar image.
(ii) **Impact of joint learning** & **feature concat.** Accuracy of all ranks (1, 5, 10) and mAP of joint learning and feature concatenation strategies is higher than the processing single representation on both datasets. On SAIVT-38 the rank-1, rank-5 and rank-10 accuracy increases by 9.86%, 6.81% and 4.76% respectively and mAP are improved by 5.05% than those results obtained by processing event-frame. Similarly, on SAIVT-58 dataset the rank-1, rank-5, rank-10 and mAP are improved by 2.98%, 4.5%, 1.02%, and 1.04%.
(iii) **Impact dissimilar camera view.** Both baseline and proposed approach accuracy drop on SAIVT-58 dataset due to the extreme camera view changes.

As this paper is the first to explore the event-driven person re-id and present benchmark results, no state-of-art frameworks are trained on event data. Hence, we trained the same network on original dataset (RGB) and called it "baseline" to observe how far is the event-driven re-id from conventional re-id in terms of accuracy. Table 2 shows that the proposed event-driven re-id did not meet the baseline performance (with conventional RGB images). Because the event camera suppresses redundant visual information and the proposed person re-id model only relies on minimum visual information (contours and edges).

DukeMTMC-reid: For DukeMTMC-reid Table 2 shows that the rank accuracy gap substantially increases between baseline and proposed method compared to the results on SAIVT. The degradation in performance is mainly due to two reasons: the acquisition setup of synthetic events

cam03 cam05 cam08

Figure 4. Reconstructed grayscale images from event-stream

generation for both datasets was different and DukeMTMC-reid is a more challenging dataset than SAIVT.

4.4. Evaluation for Privacy-Preserving

We also validate the privacy-preserving prospect of event-based person re-id, in view of the privacy attacks approaches [30, 37, 6]. Therefore, the grayscale images were reconstructed from synthetic event-stream through E2VID [30] for SAIVT dataset, shown in Fig. 4. Afterward, we train our baseline method together with state-of-the-art person re-id model BoT [20] on those reconstructed images (Gray_{event}) to examine the performance. The performance comparison of both methods trained on Gray_{event} against RGB and Gray_{rgb} (converted from RGB) plus with proposed event-based re-id are reported in Table 4, 5, and 6. Results on two subsets of SAIVT: SAIVT-38 and SAIVT-58 are shown in Table 4. and 5. while the results for the complete dataset reported in Table 6.

We notice that the performance results follow the same pattern for each method (baseline and BoT) and dataset, as can be seen in Table 4, 5, and 6. When compare Gray_{event} results against RGB and Gray_{rgb}; overall, the difference in performance is quite significant; both rank accuracy and mAP are no way near to RGB and Gray_{rgb}. Even the proposed event-based re-id outperforms the state of art re-id meodel BoT trained on Gray_{event} data. For SAIVT-38 dataset we observe in Table 4. the rank-1 accuracy difference between Gray_{event} and Gray_{rgb} is approximately 30% and between Gray_{event} and RGB is 33%. Similarly, for SAIVT-58 the accuracy gap is 33.5% and 36.4% as reported in in Table 5. Lastly, in Table 6. for the complete dataset which comprises of eight cameras, the difference is 29.8% 34.6% in the same manner. Despite the fact that the grayscale images can be recover from event streams, the significantly lower performance of person re-id on the recovered data justify and strengthen the presumption of event-driven privacy-preserving person re-id.

4.5. Ablation Studies

To further analyze the contribution of polar feature loss (l_p) as introduced in (6), we perform an ablation study shown in Table 3. Initially, the polar loss l_p is shown to be vital to our re-id network because we notice rank-1 accuracy drop 9.86 % and mAP 5.05 % on SAIVT-38 when the loss was excluded. This is caused by no direct supervision to guide our re-id model to learn discriminant features from edge patterns in the polar images, and thus the resulting model suffers from edges information loss. By introducing the polar loss ($\beta = 1$) the performance of our re-id model improved. However, we further investigate to find the optimal value of the coefficient β, which controls the contribution of the loss. In Table 3, for $\beta = 0.3$ we can observe that the rank1 accuracy and mAP additionally increased by 6.8% and 3.91% respectively with comparison to $\beta = 1$.

Table 4. Person re-id performance comparison on reconstructed grayscale images vs standard RGB/grayscale images of SAIVT-38 dataset in terms of rank accuracy and mAP.

Method	Data	R1	R5	R10	mAP
BoT[20]	RGB	90.8	96.9	97.9	85.6
	Gray_{rgb}	87.8	93.8	96.9	80.4
	Gray_{event}	57.9	68.7	75.6	49.9
$\text{Our}_{Baseline}$	RGB	87.8	93.9	97.0	82.8
	Gray_{rgb}	84.7	91.8	93.9	78.2
	Gray_{event}	58.2	69.4	77.6	51.7
$\text{Our}_{(Event+Polar)}$		*73.5*	*84.7*	*89.8*	*55.3*

Table 5. Person re-id performance comparison on reconstructed grayscale images vs standard RGB/grayscale images of SAIVT-58 dataset in terms of rank accuracy and mAP.

Method	Data	R1	R5	R10	mAP
BoT[20]	RGB	55.9	67.7	75.5	52.9
	Gray_{rgb}	53.0	65.9	71.8	50.1
	Gray_{event}	19.5	31.6	39.9	14.4
$\text{Our}_{Baseline}$	RGB	53.9	65.7	72.6	51.2
	Gray_{rgb}	51.8	64.7	70.6	49.6
	Gray_{event}	19.7	30.3	40.1	14.6
$\text{Our}_{(Event+Polar)}$		*24.6*	*37.8*	*47.1*	*15.5*

Table 6. Person re-id performance comparison on reconstructed grayscale images vs standard RGB/grayscale images of SAIVT dataset in terms of rank accuracy and mAP.

Method	Data	R1	R5	R10	mAP
BoT[20]	RGB	75.9	84.7	87.8	54.9
	Gray_{rgb}	71.1	80.9	85.4	44.1
	Gray_{event}	41.3	58.7	64.2	21.3
$\text{Our}_{Baseline}$	RGB	74.2	83.3	86.2	53.1
	Gray_{rgb}	69.4	79.8	84.6	42.4
	Gray_{event}	40.7	57.6	63.4	20.1
$\text{Our}_{(Event+Polar)}$		*49.8*	*63.4*	*70.7*	*25.8*

4.6. Challenges and Limitations

Despite the advantages of event cameras in person re-id it has been held back by the unavailability of event-based

Table 7. Ablation study of the loss function on SAIVT-38 dataset. Note that, each row indicates the different value of coefficient β

Method	R1	R5	R10	mAP
L_T=$\alpha L_e + \beta L_p$ w/ α =1				
$\beta = 0$	63.61	77.89	85.04	50.22
$\beta = 0.1$	65.31	78.57	86.73	52.98
$\beta = 0.2$	67.35	81.63	86.73	49.64
$\boldsymbol{\beta = 0.3}$	**73.47**	**84.70**	**89.80**	**55.27**
$\beta = 0.4$	68.37	83.67	89.80	53.15
$\beta = 0.6$	63.27	74.50	85.71	47.01
$\beta = 0.8$	63.61	77.89	85.04	50.22
$\beta = 1$	66.67	81.29	87.08	51.36

datasets. To create synthetic dataset with event-simulator it required to process image sequences data. Very few person re-id datasets are available with full-frame sequences (tracklets); mostly are either bounding boxes with tracklets or full-frame without tracklets. In contrast with SAIVT which is a video-based dataset, we applied an event camera model to DukeMTMC-reid dataset, homegraphic movements on the image plane to generate event-data. But it also produces events for the background region which accumulate noise in event-frames see Fig. 5. To address these difficulties and for future research dataset captured with event camera would be required.

All three dataset (SAIVT-38, SAIVT-58, and DukeMTMC-reid) poses different re-id challenges. SAIVT-38 suffers from illumination and background variation while SAIVT-58 includes camera view variation; additionally, DukeMTMC-reid comprises occlusion and background clutter. With increasing level of difficulties from SAIVT-38→SAIVT-58→DukeMTMC-reid Table 1 and Table 3 shows the performance also degrade in the similar way. We can conclude that the main challenges in event-driven re-id are viewpoint variation, occlusion and background clutter. These challenges are also considered in classical person re-id.

(a) (b)

Figure 5. synthetic events includes background noise which can be noted in (a) & (b) event-frame

5. Conclusion and Discussion

In this paper, we proposed to solve the person re-id problem in event-based vision. Since event cameras capture scene dynamics without providing RGB image content, event-frames deliver mostly edge and texture contours details that might be used for privacy-preserving re-id. In the following, we discuss the main findings of this work.

The proposed method, even if using the minimal visual information given by event cameras, shows that in specific setups, we are still able to achieve event-driven re-id. However, the proposed approach did not reach the accuracy of conventional re-id (based on standard RGB images). Experimental results on two synthetic datasets suggested that event-driven person re-id can tackle illumination and background variation challenges, but they struggle to deal with occlusions and pose/view changes.

This paper indeed shows that event-driven re-id approach has several challenges that remain ahead; future work can be structured to tackle such issues. In particular, to construct an event-based re-id benchmark, capture with event sensors. Subsequent efforts should be posed to reduce the impact of viewpoints changes, possibly by leveraging different viewpoints in synthetic scenarios as provided by recent dataset [10]. As occlusions reduce performance, having approaches that can detect human body parts from event-frames might support inference [33]. In this way, it might be possible to associate the events related to real body parts of the target and not with other occluding bodies.

Considering existing machine learning techniques that reconstruct images from event-stream might constrain event-based privacy-preserving person re-id. This work also evaluates the efficacy of privacy-preserving re-id using recovered grayscale data. The accuracy of the state-of-the-art re-id model trained on these recovered images is substantially lower than that trained on standard RGB or grayscale images. That Implies event-based sensor modality is a step towards privacy-preserving person re-id. However, this issue should be investigated further before pushing forward event-sensor as a complete privacy-compliant solution. Note that the security aspect of event sensors has already been addressed in [9] that uses a dedicated encryption framework for event stream. Indeed these approaches are necessary to provide a robust privacy-preserving solution using event-cameras networks, and they should be adopted for further research to assure the safe deployability of such solutions.

References

[1] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Com-*

puter Vision and Pattern Recognition (CVPR), pages 7243–7252, 2017.

[2] I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *European Conference on Computer Vision*, 2012.

[3] Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *WACV*, 2016.

[4] R. Benosman, C. Clercq, X. Lagorce, S. Ieng, and C. Bartolozzi. Event-based visual flow. *IEEE transactions on neural networks and learning systems*, 25(2):407–417, 2013.

[5] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, pages 1–8, 2012.

[6] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *CVPR*, 2020.

[7] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10257–10266, 2020.

[8] Julia Dietlmeier, Joseph Antony, Kevin McGuinness, and Noel E O'Connor. How important are faces for person re-identification? In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021.

[9] Bowen Du, Weiqi Li, Zeju Wang, Manxin Xu, Tianchen Gao, Jiajie Li, and Hongkai Wen. Event encryption for neuromorphic vision sensors: Framework, algorithm, and evaluation. *Sensors*, 2021.

[10] M. Fabbri, G. Braso, G. Maugeri, A. Osep, Riccardo Gasparini, Orcun Cetintas, Simone Calderara, Laura Leal-Taixe, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *International Conference on Computer Vision (ICCV)*, 2021.

[11] Alem Fitwi and Yu Chen. Privacy-preserving selective video surveillance. In *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, 2020.

[12] Alem Fitwi, Yu Chen, Sencun Zhu, Erik Blasch, and Genshe Chen. Privacy-preserving surveillance as an edge service based on lightweight video protection schemes using face de-identification and window masking. *Electronics*, 2021.

[13] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3595, 2020.

[14] D. Gehrig, A. Loquercio, K. G Derpanis, and D. Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5633–5643, 2019.

[15] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza. Eklt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3):601–618, 2020.

[16] J. Guo, Y. Yuan, L. Huang, C. Zhang, J. Yao, and K. Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3642–3651, 2019.

[17] M. S. Hanif, S. Ahmad, and K. Khurshid. On the improvement of foreground–background model-based object tracker. *IET Computer Vision*, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] X. Lagorce, G. Orchard, F. Galluppi, B. Shi, and R. Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[20] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[21] A. I Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5419–5427, 2018.

[22] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos. Event-based moving object detection and tracking. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9, 2018.

[23] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017.

[24] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 2018.

[25] G. Orchard, C. Meyer, R. Etienne-Cummings, C. Posch, N. Thakor, and R. Benosman. Hfirst: a temporal approach to object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2028–2040, 2015.

[26] L. Pang, Y. Wang, Y. Song, T. Huang, and Y. Tian. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 609–617, 2018.

[27] Marina Paolanti, Luca Romeo, Daniele Liciotti, Rocco Pietrini, Annalisa Cenci, Emanuele Frontoni, and Primo Zingaretti. Person re-identification with rgb-d camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection. *Sensors*, 2018.

[28] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *CVPR*, 2021.

[29] H. Rebecq, D. Gehrig, and D. Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982, 2018.

[30] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, 2019.

[31] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[32] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[33] G. Scarpellini, P. Morerio, and A. Del Bue. Lifting monocular events to 3d human poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[34] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018.

[35] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7244–7253, 2019.

[36] G. Wang, J. Lai, P. Huang, and X. Xie. Spatial-temporal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8933–8940, 2019.

[37] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In *CVPR*, 2020.

[38] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[39] Kok-Seng Wong, Nguyen Anh Tu, Anuar Maratkhan, and M Fatih Demirci. A privacy-preserving framework for surveillance systems. In *2020 the 10th International Conference on Communication and Network Security*, 2020.

[40] Q. Yang, A. Wu, and W. Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[41] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[42] Alex Z., Liangzhe Y., Kenneth C., and Kostas D. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Proceedings of Robotics: Science and Systems*, 2018.

[43] X. Zhu, P. Morerio, and V. Murino. Unsupervised domain-adaptive person re-identification based on attributes. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2019.