IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 March 2024, accepted 6 May 2024, date of publication 10 May 2024, date of current version 17 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399539

RESEARCH ARTICLE

Event Anonymization: Privacy-Preserving Person Re-Identification and Pose Estimation in Event-Based Vision

SHAFIQ AHMAD[1,2], PIETRO MORERIO[1], (Member, IEEE), AND ALESSIO DEL BUE[1], (Member, IEEE)

[1]Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia at Italy (IIT), 16152 Genoa, Italy
[2]Università degli Studi di Genova, 16145 Genoa, Italy

Corresponding authors: Shafiq Ahmad (shafiq.ahmad@iit.it), Pietro Morerio (pietro.morerio@iit.it), and Alessio Del Bue (alessio.delbue@iit.it)

ABSTRACT The widespread use of visual surveillance in public areas puts individual privacy at stake while also increasing resource usage (energy, bandwidth, and computation). Neuromorphic vision sensors (or event cameras) are considered viable solutions for privacy issues; since event cameras only capture scene dynamics, they do not capture detailed RGB images of individuals. However, recent deep learning architectures have enabled the reconstruction of high-fidelity images from event sensor data that could reveal individual identity information. As a result, it reintroduces privacy risks for event-based vision applications. In this work, we focus on protecting the identity of individuals from such image reconstruction attacks by anonymizing event data. To achieve this, we present an end-to-end network architecture jointly optimized for the twofold objective of preserving privacy and performing a downstream computer vision task. The proposed network learns to scramble events, thereby degrading the quality of images that potential intruders could reconstruct. We demonstrate the application of our framework in two challenging computer vision tasks: person re-identification (ReId) and human pose estimation (HPE). To this end, we also present and evaluate the first event-based person ReId dataset, Event-ReId. We validate the privacy-preserving efficacy of our approach against possible privacy attacks through extensive experiments: for person ReId, we utilize the real event-based Event-ReId dataset and synthetic event data simulated from the SoftBio dataset; for HPE, we use a publicly available event-based dataset DHP19. The results of both tasks show that anonymizing event data effectively protects private information with minimal impact on the subsequent task performance.

INDEX TERMS Neuromorphic vision, event camera, event anonymization, privacy-preserving, person re-identification, human pose estimation.

I. INTRODUCTION

Intelligent surveillance systems used for security and monitoring are installed in both personal spaces (e.g., home surveillance) and public urban areas (including hospitals, banks, shopping malls, airports, and streets). While these always-connected vision sensors are useful, they bring up several concerns: 1) ethical debates over balancing safety and security needs against individual privacy rights; 2) the risk of unauthorized access to sensory data, which could compromise user privacy; 3) the high resource demands of extensive sensor networks, such as energy, bandwidth, and computing power. A notable advancement in this field is the adoption of neuromorphic vision sensors, also known as event cameras. These sensors differ from conventional RGB cameras because they capture only brightness changes in the scene, not detailed visual images of people, thus offering a degree of privacy by design. Additionally, their ultra-low resource usage makes them highly suitable for continuous operation. Finally, their high dynamic range allows them to function effectively in various lighting conditions, including those that are typically challenging. Similar to conventional

The associate editor coordinating the review of this manuscript and approving it for publication was Yangmin Li.

RGB cameras, event cameras are capable of performing a range of vision tasks. These include object recognition [1], human pose estimation [2], [3], detection and tracking [4], [5], [6], and person re-identification (ReId) [7].

Event cameras output asynchronous events triggered with ultra-low latency when the brightness changes at the pixel level surpass a given threshold. The nature of the event stream is asynchronous; therefore, it does not create conventional images. Instead, it generates a data stream composed of activated pixel positions (i.e., (u, v) coordinates) and their polarity at a given timestamp. These event streams were considered privacy-preserving [7] due to their lack of capturing detailed visual information that would enable the recognition of individual characteristics, such as faces, by either humans or algorithms. However, event streams encode the entire visual signal in highly compressed form and could, in principle, be decompressed to retrieve a high-quality video stream. Recently, deep neural network-based event-to-image conversion models [8], [9], [10], [11] have demonstrated impressive abilities in recovering gray-scale images from event streams, posing a potential risk to the privacy aspect of event-based vision applications. As illustrated in Figure 1(a), event-to-image [8] model reconstruct a detailed image that reveals personal identity information; consequently, this implies that event cameras can no longer be deemed inherently privacy-preserving.

To counteract privacy attacks (event-to-image) on event data, Du et al. [12] recently introduced a manually designed encryption framework to counteract privacy attacks on event data. This method uses spatial chaotic mapping to scramble event positions and invert their polarities. Consequently, the spatial information in the encrypted event stream becomes distorted due to the 2D position scrambling, and event-to-image conversion techniques are ineffective in producing high-quality images. However, the limitation of this encryption approach is the incompatibility with the direct execution of downstream computer vision tasks. The encrypted event stream primarily prevents privacy attacks during transmission or storage, requiring decryption prior to any practical application.

In this work, we introduce a learning-based privacy-preserving strategy known as event-stream anonymization. This approach aims to prevent event-to-image methods from converting event data to high-quality images that may contain privacy information (refer to Figure 1 (b)) while still facilitating the execution of downstream tasks, such as person ReId and human pose estimation (HPE). The event anonymization approach focuses explicitly on degrading the quality of images that a privacy attacker might reconstruct (i.e., event-to-image module [8]) while jointly optimizing a downstream task (e.g., ReId, HPE) through an end-to-end manner.

In general, a privacy-preserving model should preserve key visual privacy information, including identity (such as faces), gender, race, color, etc [13], [14]. However, the definition of privacy information for a privacy-preserving model may differ based on its application. In principle, the main objective of our privacy-preserving model within the event-based vision system is to prevent image reconstruction attacks and anonymize personal identity information. In other words, the proposed framework ensures that individuals remain unrecognizable in the resultant gray-scale images even if an attacker attempts to reconstruct images from an anonymized event stream. An example scenario involves an attacker using a person's name and photo and aiming to identify that person by maliciously accessing the event camera network with the intent of reconstructing image data. The proposed anonymization framework averts this privacy invasion yet still enables downstream tasks of ReId and HPE by the surveillance system. For instance, the two tasks, anonymization and ReId, seem to have contrasting objectives. This represents the actual challenge of our work. In any case, the person ReId only aims to associate images of a person in the event camera network, whereas anonymization is about protecting a person's identity or other biometric traits.

The primary goal is to achieve privacy-preserving person ReId and HPE in event-based vision; thus, we evaluate the performance of event anonymization on downstream tasks of ReId and HPE along with measuring its ability to protect privacy. The event anonymization framework's potential in privacy protection is tested in two ways: (i) first, the robustness against image reconstruction attack (event-to-image) is measured by the (poor) quality of the reconstructed gray-scale images; (ii) second, anonymizing identity information is assessed by verifying that person identification through classic full-body or face recognition is hardly possible using reconstructed images. Moreover, the extraction of gender attributes from reconstructed gray-scale images is regarded as sensitive private information. We also quantify whether the attacker could attempt to determine an individual's gender (male or female) from reconstructed images. Finally, the most effective privacy protection is assured when no adversary can learn or recover the privacy information by attacking the privacy-preserving model. Thus, we test event anonymization robustness to *inversion attack*, where an attacker attempts to reverse the anonymization (see Figure 1 (d)). In another scenario, if the attacker retrains the image reconstruction module to infer high-quality images from the anonymized data, this threat is known as *adversarial learning*. The proposed approach effectively anonymizes the event stream with minimal impact on person ReId performance accuracy, as shown by comprehensive tests on simulated event data and real event-based person ReId dataset, Event-ReId. Similarly, the proposed approach also achieves analogous results for human pose estimation.

This paper is the extension of our earlier work [15], which demonstrates the preliminary results of the proposed method for person ReId task only by utilizing a relatively small event-based ReId dataset. In contrast to the [15], this work not only focuses on a downstream task of person ReId but also investigates a human pose estimation task as well; we perform

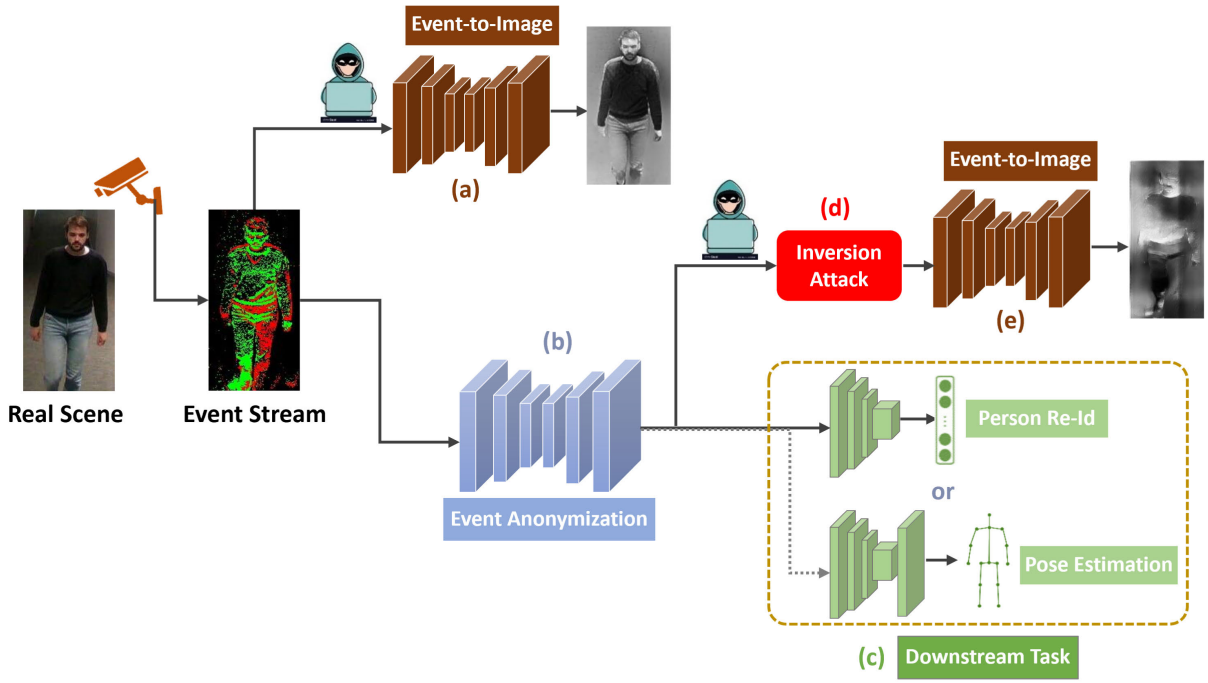


FIGURE 1. Converting an event stream into a high-quality gray-scale image, i.e., Event-to-Image [8], is regarded as a Privacy Attack, which could reveal an individual's personal identity information. (a). In response, we have developed a novel network architecture for Event Anonymization (b) to counteract such privacy threats by scrambling the event stream. This scrambling ensures that reconstruction quality degrades while the effectiveness of the event-based applications, e.g., person ReId or human pose estimation (c), remains unaffected. Furthermore, we explore the scenario of an Inversion Attack (d), where an adversary might try to reverse our anonymization process to reconstruct a high-quality image (e).

the following new analysis, bringing in new contributions to the relevant domains.

- We extend the event anonymization approach to another computer vision application, such as human pose estimation (HPE). Integrating the event anonymization framework into HPE tasks ensures maintaining individual privacy in applications like activity recognition, where pose data is crucial but should not contain personal identity information. We implement the same pipeline of event anonymization [15] but with different downstream applications of HPE.
- While the initial experiments on the small event-based ReId dataset yielded promising results, evaluating the event-based person ReId approach on a larger dataset is essential. Hence, we captured additional data for event-ReId, increasing the number of identities from 33 IDs to 60 IDs, and we named the extended version of the dataset ***Event-ReId-v2***. With this newly acquired, larger dataset, we repeated all the experiments from [15] and reported the updated results.
- We analyze the privacy aspects of the event anonymization model against a wider range of potential privacy threats. This includes gender classification and adversarial learning attacks, in addition to those explored in [15].

II. RELATED WORK

This section first reviews the current available real and synthetic event-based datasets. Then, presents a review of the recent privacy-preserving approaches in standard (RGB) and event-based vision sensors. Finally, we discussed reviews of the privacy-preserving person ReId approaches and event-based person ReId and HPE methods.

A. EVENT CAMERA DATASETS

Due to the relatively new vision-sensing technology, only a few event-based datasets that are captured with an event camera are available. Among these, human pose estimation [2] dataset, action recognition [16], [17], [18] dataset, face expression recognition [19] dataset, and car recognition [20], [21] datasets. To address the limited availability of event data, researchers have alternatively suggested the generation of semi-synthetic and synthetic event-based datasets. This initiative aims to stimulate new research utilizing event cameras for various tasks, as highlighted in Rebecq et al. work [22].

Semi-synthetic datasets such as those in [23] and [24] convert standard video into event data. For instance, [23] transforms VOT2015 and UFC50 video datasets into Dynamic Vision Sensor (DVS) data by recording them from a 60 Hz LCD monitor using a DVS camera. Similarly, [24] creates event-based versions of MNIST and Caltech101 by displaying their frames on a screen and capturing the data with an event camera mounted on a pan-tilt motor. Moreover, generating event-based synthetic datasets can be facilitated using event camera simulators [22], [25], [26]. These simulators are software tools designed to emulate the

functionality of physical event cameras, generating synthetic event data that can be used for research and development purposes.

B. PRIVACY-PRESERVING IN STANDARD (RGB) VISION

Several techniques have been formulated to address privacy protection challenges in standard RGB cameras. These approaches [14], [27], [28], [29], [30] fall into two categories: software and hardware level defense against privacy attacks. Software-level protection methods [14], [29] use various computer vision algorithms to modify the representation of images post-acquisition. These methods utilize adversarial training to develop encodings that protect privacy, discarding sensitive visual information in image data but maintaining key features necessary for inference tasks and protecting against adversarial attacks. Conversely, the hardware-level protection framework operates directly on the vision sensor, adding an extra layer of security by eliminating sensitive data at the point of image capture. Most recent methods primarily focus on adjusting the distortion parameters of a virtual lens through adversarial training. This technique aims to obscure human identity information while still capturing the vital visual data needed for computer vision tasks, as discussed in studies [27], [28], [30]. Actual lenses can then be produced based on the parameters learned in this process.

Besides, in the current literature, privacy-preserving methods in [31], [32], [33], and [34] illustrate that federated learning offers an alternative approach for integrating privacy in computer vision systems. These methods underscore that federated learning ensures privacy by training models on-device with local data, only sharing model improvements rather than the data itself, thereby significantly reducing privacy risks associated with central data storage and processing.

C. PRIVACY-PRESERVING IN EVENT-BASED VISION

Event cameras are often considered privacy-preserving [7], [35] because they inherently eliminate detailed visual private data such as face details. However, their event stream compactly encodes the entire visual signal. Recent studies have shown that it is possible to uncompress this data and retrieve a standard (gray-scale) visual representation. This decompression has been achieved using different methods, such as patch-based dictionaries [36], variational models [37], or deep learning-based techniques [8], [9], [10], [11]. These methods of converting event data to images suggest that event cameras might not be reliably privacy-protective anymore, as attackers could train models to breach the anonymity they offer.

Du et al. [12] explore the privacy aspects of event cameras, analyzing potential threats such as the reconstruction of gray-scale images and privacy-related classifications. In addition, Du et al. [12] introduced a manually designed encryption framework that uses spatial chaotic mapping to scramble event positions and invert their polarity; as a result, event-to-image methods failed to recover images. Nevertheless, this framework is primarily effective for protecting the event stream during transmission and storage. The visual data within the event stream becomes distorted due to the 2D position scrambling; therefore, computer vision tasks (e.g., detection, tracking, person ReId, human pose estimation, etc.) can not be performed using these encrypted streams and must be decrypted before being utilized.

Contrary to other approaches, our method employs a learning-based, end-to-end strategy for privacy preservation, which scrambles event streams in a way that causes image reconstruction techniques to yield degraded images. Despite this, the method preserves the essential information to execute computer vision tasks on these scrambled events. To evaluate the fidelity of the event anonymization framework, we set our privacy-preserving model for two diverse and challenging computer vision tasks: person re-identification (ReId) and human pose estimation (HPE).

D. EVENT-BASED PERSON RE-IDENTIFICATION

Extensive research has been conducted on the person ReId problem in standard RGB camera networks, and with deep-learning-based ReId approaches [38], [39] rapidly enhancing performance. While the majority of ReId systems are designed for traditional RGB cameras, various approaches have been developed for multi-modal person ReId. These include techniques like cross-modal RGB-infrared [40], [41] and systems utilizing RGB-D cameras [42], [43].

Presently, ReId poses significant privacy challenges, making the protection of individuals' privacy a critical issue [44], especially considering the General Data Protection Regulation (EU GDPR). Only a few methods [44], [45], [46] tackle privacy issues specific to person ReId. Dietlmeier et al. [45] applied face blurring techniques to anonymize personal identities and implement the person ReId task. On the other hand, Dou et al. [44] developed a privacy-preserving approach called person identity shift (PIS), which effectively conceals the absolute identity of individuals in images while maintaining the association between image pairs. Additionally, Zhao et al. [46] introduced a cloud-based, privacy-preserving solution for person ReId. The framework enables cloud servers to execute ReId processes on encrypted data, subsequently delivering the final ReId results in an unencrypted format (in plain text).

A major drawback of all these methods is their failure to guarantee end-to-end privacy protection within the ReId system. The risk of unauthorized access to surveillance cameras remains a substantial privacy concern. Ahmad et al. [7] introduced a novel event-based person ReId system to tackle this issue. Event cameras, which record changes in a scene without producing traditional RGB images, offer a different approach. Ahmad et al. [7] demonstrated that event frames primarily provide edge and texture information, potentially useful for ReId purposes. However, as previously discussed, event-based data can still reveal personal traits by employing

deep learning methods [8], [10], [11] that can generate high-quality gray-scale images from event streams, eventually compromising privacy.

E. EVENT-BASED HUMAN POSE ESTIMATION

Human pose estimation (HPE) is considered a fundamental problem in computer vision due to a broad range of applications (e.g., motion analysis, healthcare, sports, augmented reality (AR), virtual reality (VR), autonomous driving, human-computer interaction, etc.) [47], [48], [49]. The HPE task has been exhaustively studied in traditional (RGB) vision sensors, and deep learning approaches have already produced significant advancements and impressive results [47], [50], [51], [52]. The current HPE techniques rely on standard RGB images without taking privacy into account. However, Hinojosa et al. [27] developed a privacy-preserving HPE by designing an optical lens using deep optics that hides personal information while enabling HPE. Besides, HPE has several real-time applications where low-latency pose prediction is a key feature, including gaming, accident detection, and real-time movement feedback in rehabilitation treatment. Therefore, considering the advantages of event-camera over traditional cameras, several state-of-the-art methods [2], [3], [53], [54] have been proposed to solve 2D/3D human pose estimation in event-based vision. Calabrese et al. [2] accumulate event-stream into frames to predict 2D poses from multiple views and then, through triangulation, estimate 3D pose. Scarpellini et al. [3] predict 3D pose from a single camera view with spatiotemporal voxel grids as an input instead of event frames.

III. EVENT-REID: A NEW DATASET AND BENCHMARK

Initially, we aim to build a privacy-preserving person ReId system using an event-cameras network. Yet, the research community lacks a dataset captured with real event cameras, which are also appropriate for benchmarking person ReId methods. Hence, regardless of the advantages of event cameras in a surveillance application, research has been held back by the absence of event data, and so far, only simulated experiments have been deployed [7]. To solve this problem and to facilitate new research on this topic, we captured the event-based person ReId dataset named **Event-ReId-v2** (extended/second version of the previous Event-ReId dataset [15]).

The current Event-ReId-v2 dataset comprises 60 identities walking across a disjoint field of view of four event cameras integrated through a surveillance network. The cameras are installed at various positions and angles of tilt, and each one is paired with an RGB camera in a stable stereo setup. This arrangement ensures they capture approximately the same scene and are synchronized using the network clock, see Figure 2. Each RGB camera captures data at a frame rate of 33 FPS with a resolution of 640 × 480 pixels, resulting in around 27K images. On average, each camera records 120 frames per person. The resolution of the event cameras matches that of the RGB cameras, and the duration of recording for each stream is approximately ≈4 sec, the same for both sensors, lasting around 4 seconds. Additionally, within the 60 identities captured, 9 subjects are wearing face masks. A total of 57 subjects appear in all four camera pairs, while the rest are seen in three camera views. The dataset encompasses a range of variations, including changes in lighting, poses, and viewpoints. We manually annotate the person and face bounding boxes on both event and RGB streams; the event ground truth bounding box is synchronized with RGB bounding boxes.

TABLE 1. Event-based dataset size comparison.

Dataset	**Event ReId-v2**	**Event ReId-v1 [15]**	n-HAR	DailyAction DVS	DHP19
# IDs	60	33	30	15	17
# Cam	4	4	1	2	4

Our proposed dataset demonstrates a favorable size relative to existing event-based datasets, notably in the domain of activity recognition **n-HAR** [16] and **DailyAction-DVS** [17], and human pose estimation dataset **DHP19** [2] (see comparisons in Table 1).

Note: The latest version of event-based person ReId dataset **Event-ReId-v2** can be downloaded from here **https://doi.org/10.5281/zenodo.10398002**

IV. PROPOSED METHOD

The proposed pipeline is composed of three primary components: the event anonymization block, which removes privacy-related information from the event stream; the event-to-image reconstruction block that acts as a privacy attacker; and the downstream block, which executes downstream computer vision-related tasks (such as person ReId and human pose estimation) on the anonymized event stream. In the following section, we begin by detailing the representation of input events to the network and then explain each module in-depth, detailing their implementation and functionalities for preserving privacy and downstream tasks. We conclude this section with an overview of the joint optimization method.

A. INPUT EVENT REPRESENTATION

An event camera generates an asynchronous stream of events that encodes the timestamp, location, and polarity of brightness changes (indicating an increase or decrease in intensity) [55]. Each event solely provides limited information about the appearance of the scene. Asynchronous event data are often transformed into grid-like formats, such as event frames or 2D histograms [56], time surface 2D maps [20], and voxel grids [57]. This preprocessing step aids in both the visualization of the data and the extraction of valuable information, making it compatible with standard frame-based methods, for example, deep convolutional neural networks (CNNs) [20], [56], [57].

Our network takes a voxel grid, denoted as Zhu, as its input, following the approach proposed in [57]. A voxel grid is essentially a three-dimensional space-time histogram of events created by dividing the time domain into discrete

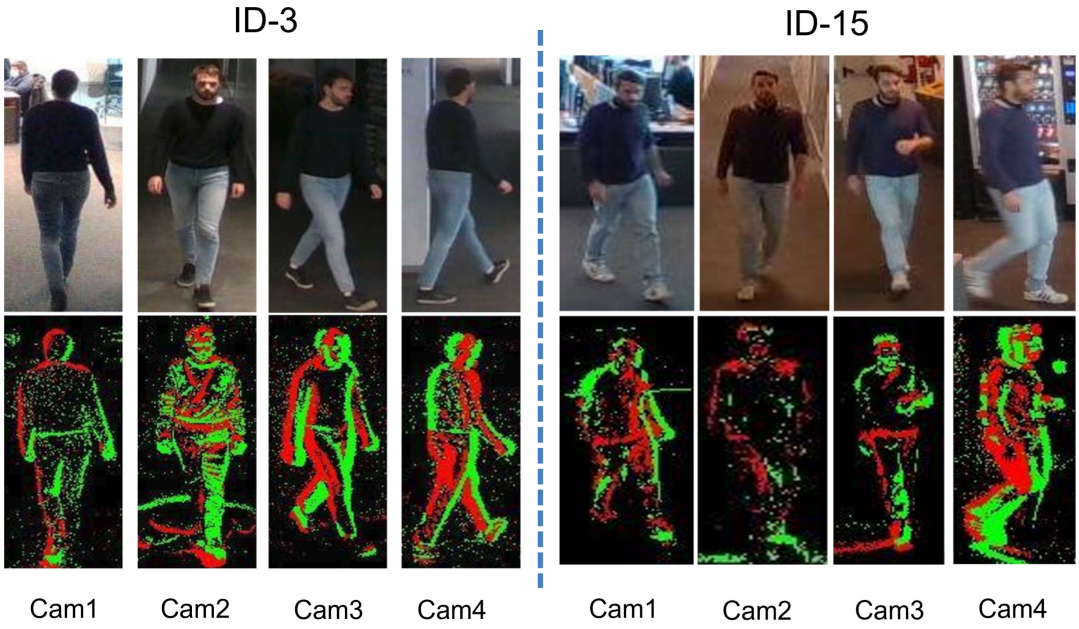


FIGURE 2. Event-ReId dataset samples: RGB and event cameras views of all four sensors.

intervals, with each voxel corresponding to a specific pixel and time period. Spatiotemporal coordinates, x_p, y_p, t_p, lie on a voxel grid such that $x_p \in \{1, 2, \ldots, W\}, y_p \in \{1, 2, \ldots, H\}$, and $t_p \in \{t_0, t_0 + \Delta t, \ldots, t_0 + B\Delta t\}$, where t_0 is the initial timestamp, Δt is the temporal bin size, and B is the number of temporal bins and W, H are the event camera width and height, respectively. We adopt the voxel grid representation for three key reasons: *i)* to ensure that the model is entirely differentiable; *ii)* because the event-to-image methods in our proposed model also depend on a voxel grid format; *iii)* a voxel grid effectively preserves the temporal information present in event streams.

B. NETWORKS AND MODULES

Event-Stream Anonymization Block: In our framework, as shown in Figure 3(a)), the anonymization network modifies the event streams. This modification is crucial to prevent image reconstruction techniques from transforming events into gray-scale images that could disclose sensitive information, such as facial features. Concurrently, this module is designed to sustain essential spatial information necessary for the effective execution of downstream tasks, e.g., person ReId and human pose estimation (HPE). The anonymization network incorporate a convolutional autoencoder [58] E_{an} which processes a raw event-voxel $X_e \in \mathbb{R}^{B\times W\times H}$ and output anonymized event-voxel $\hat{X}_e \in \mathbb{R}^{B\times W\times H}$.The adoption of an autoencoder-like architecture is primarily based on the requirement that this module should be capable of duplicating the event stream in a worst-case scenario, ensuring that the downstream task can be carried out. The autoencoder architecture comprises 4 convolutional layers, each equipped with a filter size of 3 and a stride of 1.

Image Reconstruction Block: The image reconstruction module (Figure 3(b)) consists of a pre-trained E2VID network [8] that is a recurrent neural network that reconstructs high-quality gray-scale images from the stream of events. In this block, any event-to-image method, e.g., [9], [10], [11], can be integrated as a privacy attacker. E2VID translates a continuous stream of events into a sequence of images. To achieve this, the incoming stream of events is partitioned into sequential (non-overlapping) spatiotemporal windows, each containing a fixed number of events. Similarly, we also used a fixed number of events (Sec IV-A) for the reconstruction module E_{rec}. The voxel-grid $\hat{X}_e$ is processed by E_{rec} to reconstruct the target gray-scale image. We thus encourage degradation in the recovered image to prevent identity information leakage. Note that the weights of this module are not updated during training.

Downstream Task Block

(i) Event-based Person ReId: Person ReId methods usually aim to learn a vector representation, usually a feature embedding from a CNN, of images to perform retrieval and recover images belonging to the same person Id. In our case, ReId is performed on event-stream data instead of the standard RGB signals.

We employ a ResNet-50 [59] pre-trained on ImageNet as the backbone for feature embedding (Figure 3(c) top). Unlike the event-based ReId in [7], which utilizes event-frames, our ReId module E_{reid} takes anonymized event-voxels $\hat{X}_e \in \mathbb{R}^{B\times W\times H}$ as input. We modify the original ResNet architecture to accommodate the B input channels of the voxel-grid representation and compute a 256-D feature embedding for ReId. The ReId model uses classification loss (cross-entropy) and triplet loss for all experiments and is jointly trained with the anonymization network.

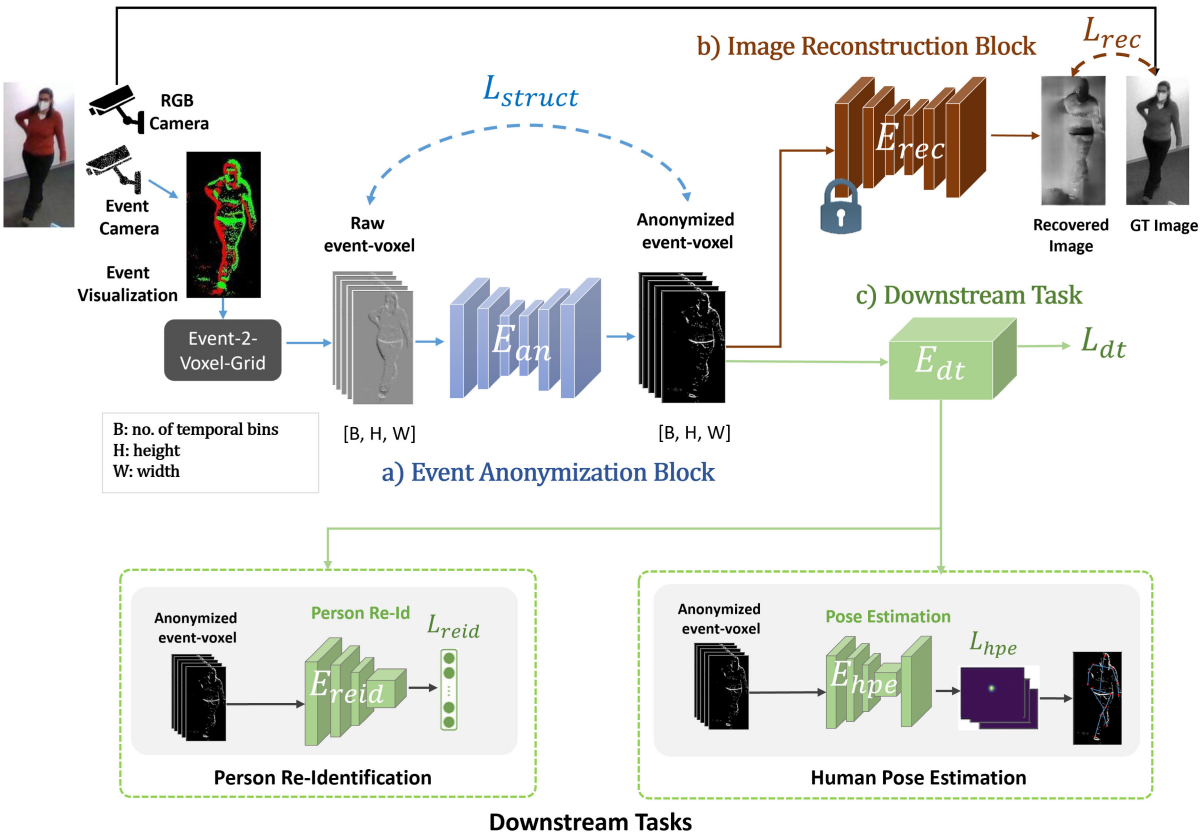


FIGURE 3. The complete pipeline of the proposed method. a) The Event Anonymization network processes raw event data (in voxel-grid form) and produces an anonymized event. A structure loss is applied to ensure the anonymization network maintains the structural information in the output anonymized voxel grid. b) The Image Reconstruction block (utilizing the pre-trained E2VID model [8]) acts as a privacy attack and tries to recover a gray-scale image. The anonymization network is designed to maximize reconstruction loss, thereby protecting personal identity information. c) The Person ReId backbone is trained on the anonymized event data, with the anonymization network in an end-to-end manner.

(ii) Event-based Human Pose Estimation: Deep learning-based pose estimation frameworks learn to predict body keypoint coordinates from input images. These approaches either directly regress the body keypoint coordinates or, via heatmap regression, estimate heatmaps that are produced by adding Gaussian kernels to each joint's position in the ground-truth [48], [60].

We design our HPE module (Figure 3(c) bottom) based on [3] and [61] approaches that combine heatmap and coordinate regression by applying soft-argmax operator on the 2D predicted heatmap to extract the normalized joints coordinates. We utilize the same backbone we use for person ReId with the same settings, except we remove ResNet-50 [59] layers after the second residual block following [3]. We process the anonymized event-voxel $\hat{X}_e \in \mathbb{R}^{B\times W\times H}$ with our HPE module E_{hpe}, first predict 2D heatmap $\hat{H}_i$ for each joint and then apply soft-argmax to get normalized joints coordinates $\hat{J}_i = (\hat{x}_i, \hat{y}_i)$:

$$\hat{H}_i = E_{hpe}(E_{an}(X_e)). \tag{1}$$

$$\hat{J}_i = soft.argmax(\hat{H}_i). \tag{2}$$

The HPE module E_{hpe} is jointly trained to predict the 2D heatmaps along with the normalized body keypoint positions; as a result, we estimate 2D human pose. For 3D pose estimation, we first estimate the 2D pose for each camera view through our trained model, and then we apply triangulation [2] to reconstruct the 3D position.

C. END-TO-END TRAINING

Our ultimate goal is to learn the parameters of anonymization network E_{an} such that: *i)* event-to-image techniques cannot recover intensity image from E_{an} output that can disclose private visual information; *ii)* downstream task achieves the best performance or at least does not experience a significant drop compared to using a non-anonymized event stream. The three modules are combined as shown in Figure 3 so that the output of E_{an} (anonymized stream) is the input of E_{rec} and E_{dt} at once. We jointly train all the E_{an} and E_{dt} modules in an end-to-end manner; each loss is described in detail below.

E_{an} has the aim of neutralizing the reconstruction attack; thus, ultimately, it must be trained with the objective of degrading the quality of the reconstructed images I_{rec}:

$$I_{rec} = E_{rec}(E_{an}(X_e)). \tag{3}$$

To evaluate the quality of the reconstructed image I_{rec} in comparison to the ground-truth image I_{gt}, we employ the structural similarity index (SSIM) [62].

SSIM [58], [62] is a metric used to measure the similarity between two images. It is based on the perception that the human visual system is highly adapted for extracting structural information from a visual scene. Thus, SSIM considers changes in texture, luminance, and contrast rather than just analyzing the pixel-by-pixel differences like traditional metrics such as Mean Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR).

$$\text{SSIM}_{(I_{rec},I_{gt})} = \frac{(2\mu_{I_{rec}}\mu_{I_{gt}} + C_1)(2\sigma_{I_{rec}I_{gt}} + C_2)}{(\mu_{I_{rec}}^2 + \mu_{I_{gt}}^2 + C_1)(\sigma_{I_{rec}}^2 + \sigma_{I_{gt}}^2 + C_2)} \quad (4)$$

where:

- I_{rec} and I_{gt} are reconstructed and ground truth images, respectively, being compared.
- $\mu_{I_{rec}}$ and $\mu_{I_{gt}}$ are the average pixel values of images I_{rec} and I_{gt}, respectively.
- $\sigma_{I_{rec}}^2$ and $\sigma_{I_{gt}}^2$ are the variances of I_{rec} and I_{gt}.
- $\sigma_{I_{rec}I_{gt}}$ is the covariance of I_{rec} and I_{gt}.
- C_1 and C_2 are constants that stabilize the division with a weak denominator.

Our objective is to degrade the quality of the recovered image; thus, we use the SSIM function, which is bounded by ranges from 0 to 1. A value close to 0 in this range signifies a lower similarity between the two images being compared.

Reconstruction Loss: Hence, during training, the $\mathcal{L}_{rec}$ loss is maximized to ensure that the images reconstructed by the attacker differ significantly from the real ones.

$$\mathcal{L}_{rec} = 1 - SSIM(I_{rec}, I_{gt}). \quad (5)$$

Structural Loss: Additionally, since our anonymization model transforms the raw event voxel, that might lead to the loss of useful visual information within the event voxel, which could significantly reduce the efficacy of downstream tasks. To maintain the structural similarity between X_e and $\hat{X}_e$, that is crucial for downstream tasks; we calculate the structural loss as follows:

$$\mathcal{L}_{struct} = 1 - SSIM(\hat{X}_e, X_e), \quad (6)$$

in this case, the $\mathcal{L}_{struct}$ is minimized to preserve structure information in output anonymized event voxel-grid.

Downstream Task Loss: Since we are performing two different downstream tasks; therefore, we trained our proposed network separately for each task with the objective functions $\mathcal{L}_{dt}^{reid}$ and $\mathcal{L}_{dt}^{hpe}$ for ReId and HPE respectively.

ReId Loss: To implement person ReId, we adopt softmax loss for classification and triplet loss for metric learning, two commonly used loss functions in several deep ReId approaches [39], [63]. The person ReId task is considered a multi-class classification problem for basic discrimination learning [39]; therefore, each pedestrian is treated as a different class and uses their IDs as a classification label to train the deep neural network. Hence, classification (or cross-entropy) loss is alternatively called identity loss:

$$\mathcal{L}_{\text{id}} = -\sum_{i=1}^{N} q_i \log(p_i) \quad (7)$$

where N is the number of sample identity categories or classes, q_i represents the ground truth probability (1 for the correct class and 0 for all other classes), and p_i is the predicted probability for class i.

In addition, triplet loss is one of the most popular depths metric losses in person ReId [63] that further improves the discriminative property by increasing the inter-class discrepancy while decreasing intra-class distinctness:

$$\mathcal{L}_{\text{triplet}} = \max\left(D_{ap} - D_{an} + \alpha, 0\right) \quad (8)$$

where D_{ap} is the distance between the anchor and the positive sample in the embedded space, D_{an} denotes the distance between the anchor and the negative sample in the embedded space, and α denotes the margin between the two distances.

We use identities labeled information from training data and apply cross-entropy loss as identity loss $\mathcal{L}_{id}$ to the output feature vector $E_{reid}(\hat{X}_e)$ and also triplet loss $\mathcal{L}_{triplet}$. As a result, the final ReId loss function can be formulated as:

$$\mathcal{L}_{dt}^{reid} = \mathcal{L}_{id}(Q_{id}, (E_{reid}(\hat{X}_e))) + \mathcal{L}_{triplet}(E_{reid}(\hat{X}_e)) \quad (9)$$

where Q_{id} is the ids label for person ReId.

HPE Loss: During HPE training, the loss is calculated between the predicted heatmap and synthetic ground-truth heatmap (generated through 2D spherical Gaussian mean-centered on the ground-truth body key points). Typically, we can train the HPE network by minimizing MSE loss between the output heatmap and target heatmap [48]; however, Jensen-Shannon divergence (JSD) based loss combined with a geometrical loss between predicted joints and ground-truth has proven to be effective [3], [61]. Similarly, we apply JSD between predicted H and ground truth $\hat{H}$ heatmaps (equation 5) and also geometric loss between predicted joints and ground truths joints.

$$D_{\text{JS}}(H||\hat{H}) = \frac{1}{2}D_{\text{KL}}(H||\hat{H}) + \frac{1}{2}D_{\text{KL}}(\hat{H}||H) \quad (10)$$

The final HPE loss is a combination of the D_{JS} and geometric loss ($\mathcal{L}_{geometric} = ||\hat{J} - J||_2$):

$$\mathcal{L}_{dt}^{hpe} = ||\hat{J} - J||_2 + D_{\text{JS}}(H||\hat{H}) \quad (11)$$

Hence, our training approach simultaneously incorporates event-stream anonymization and downstream tasks during the training phase, and the total cost function can be expressed as follows:

$$\mathcal{L}_{Total} = \alpha\mathcal{L}_{struct} + \beta\mathcal{L}_{rec} + \gamma\mathcal{L}_{dt}. \quad (12)$$

V. EXPERIMENTAL DETAILS AND EVALUATION

In this section, we show the potential of our proposed model as a method of privacy-preserving in the event-based vision for two computer vision tasks, such as person ReId and pose estimation with anonymized event voxel-grid inputs. In Section V-A, we explained the evaluation metrics for privacy-preserving and downstream tasks. Then, from Section V-B to V-E, we presented event-based privacy-preserving person ReId and HPE and their results in detail. We additionally evaluate the efficacy of the events anonymization network against three possible privacy attacks, gender prediction, inversion attack, and adversarial learning, explained in Sections V-F, V-G, and V-H, respectively.

A. EVALUATION METHODS

In order to evaluate the effectiveness of our complete model in both image reconstruction and mitigating possible privacy invasions, we must consider the trade-off between accomplishing downstream computer vision tasks (like person ReId and HPE) and maintaining privacy. Our approach involves first passing the raw event stream through the anonymization network during the inference stage to generate anonymized event data. We then evaluate the downstream task and privacy-preserving performance using the anonymized data.

Downstream Task (ReId & HPE) Evaluation: Our main goal is to perform computer vision tasks (e.g., person ReId and HPE) with anonymized event data without compromising performance accuracy. Thus, we train our downstream task backbone on anonymized and raw events separately and then compare their performance. We report the rank accuracy and mean average precision for real and simulated data for the person ReId task. While for HPE, we compute 2D and 3D Mean Per-Joint Precision Error (MPJPE).

Privacy-Preserving Evaluation: Consider the case in which the attacker can access the anonymized event data and tries to disclose the person's identity by employing image reconstruction, e.g., E2VID [8]. To experimentally test the robustness of our event stream anonymization approach against the reconstruction attack, we measure the image quality using the structural similarity index (SSIM) and peak-signal-to-noise ratio (PSNR). Low values of SSIM and PSNR suggest low image quality, which is what we expect to achieve if anonymization is successful. We compute the average SSIM and PSNR for all images in test sets of the real and simulated datasets.

In addition, we also validate that our proposed identity anonymization framework completely removes information that can be used to identify the persons. Therefore, we also formulate the privacy attack as an image retrieval and face verification task.

(i) Image Retrieval: We consider that an attacker has access to the event-based privacy-preserving surveillance camera network and holds a query image of a target to identify. The query image is either captured with a standard RGB camera Q_{rgb} or a gray-scale image $Q_{noprivacy}$ reconstructed from an event stream without the protection of the privacy module. Then, the attacker determines whether this person exists in the gallery set $G_{privacy}$ that contains degraded images by using the query image to retrieve the correct target identity.

Higher retrieval performance indicates a lower privacy-preserving effect: E_{an} performance is evaluated based on the rank accuracy or mean average precision metrics. For this experiment, we employ the state-of-the-art person ReId model BOT [64] to evaluate image retrieval and use the test sets of real and simulated datasets.

(ii) Face Recognition: In this experiment, we assume a similar scenario, where the attacker holds a *face* image (RGB or reconstructed gray-scale image) and tries to disclose identity information by matching it with a degraded face image. We use the pre-trained face recognition model ArcFace [65] to measure the resilience of our system to this privacy attack. We measure face recognition performance in terms of the area under the curve (AUC) of the ROC curve.

B. PERSON REID EXPERIMENTAL SETUP AND DATASETS

Datasets: We test our proposed event-anonymization method for the person ReId task using synthetic event data and the real event data presented in Section III. Synthetic event data is generated from the video-based person ReId SoftBio [66] dataset using the open-source event simulator [26]. The SoftBio dataset comprises 152 identities and 64,472 frames collected with eight surveillance cameras. The dataset is recorded in an uncontrolled environment, and each identity may only appear in a subset of cameras, which collect data under very different viewpoints, with drastic changes in illumination and background. In addition, we benchmark our approach on the **Event ReId-v2** dataset described in Sec. III.

Person ReId Evaluation Metrics: To evaluate the efficacy of person ReId, we used both cumulative matching characteristics (CMC) and mean average precision (mAP). The adoption of CMC and mAP has become predominant in the domain because a single evaluation metric is often insufficient to assess the overall effectiveness of the person ReId. CMC-k, or Rank-k matching accuracy, quantitatively assesses the probability of a correct match being present within the top-k retrieved results. This metric demonstrates precision in scenarios where each query is associated with a single ground truth, focusing exclusively on the first match in the evaluation process. On the other hand, the mAP metric evaluates average retrieval performance when dealing with multiple ground truths. In the field of ReId assessment, mAP successfully handles cases in which two systems perform identically in identifying the initial ground truth but differ in their capacity to recover further difficult matches.

Experimantal Setup: We generate simulated event data from SoftBio, splitting 152 identities randomly, with 76 IDs for training and another 76 IDs for testing. In the case of real

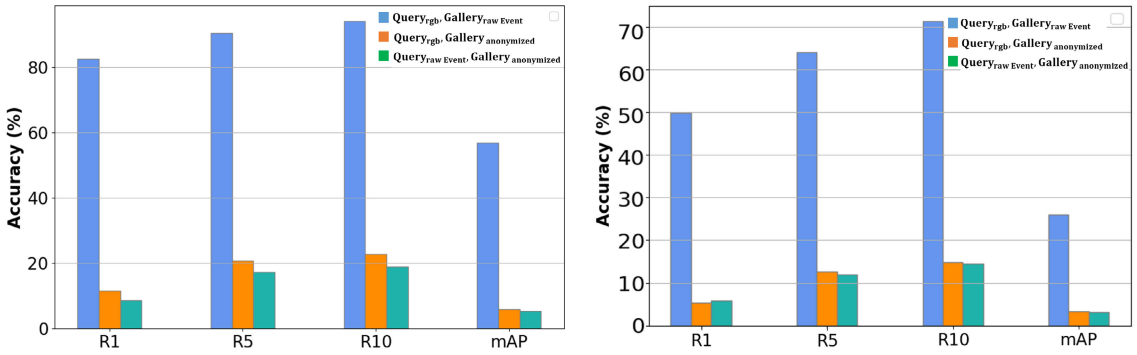


FIGURE 4. Image retrieval evaluation on Event-ReId dataset (left) and SoftBio dataset (right), following query-gallery setting, blue: query[rgb], gallery[no privacy], orange: query[no privacy], gallery[privacy], green: query[rgb], gallery[privacy].

data for Event-ReId-v2, out of 60 identities, we randomly choose 45 IDs for training and the remaining 15 IDs for testing.

We choose the time duration for the spatiotemporal voxel grid T≈40ms for synthetic event data and T≈30ms for real event data to be synchronized with the corresponding RGB frames. We set the size of temporal bin $B = 5$ for the event voxel grid, Following [8], and during training, our model resized the event voxel grid to $5 \times 392 \times 192$. We use a batch size of 32 and train the model with a base learning rate of 0.001 for 60 epochs. We set momentum $\mu = 0.9$ and the weight decay to 5×10^{-4}. In equation 4 we set $\alpha = \gamma = 1$ and $\beta = -1$. The implementation is based on PyTorch.

C. PERSON ReId RESULTS

Privacy-preserving performance: We present the image retrieval performance score for the real dataset Event-ReId and similarly for the simulated event data of SoftBio in Figure 4. The testing approaches "$Q_{rgb} \Leftrightarrow G_{privacy}$" and "$Q_{no-privacy} \Leftrightarrow G_{privacy}$" measure the retrieval score on the anonymized (privacy-preserving) image gallery using original RGB and recovered gray-scale query images respectively. For comparison, the testing approach "$Q_{rgb} \Leftrightarrow G_{no-privacy}$" measures the image retrieval score on original gallery images. The tested retrieval model BoT [64] did not perform well on our anonymized images, and for both datasets, the retrieval score is random.

Regarding the face recognition performance, Figure 5 shows the ROC curves for each testing approach: ***RGB*** measures the face verification score on original RGB face images; ***No Privacy*** measure face verification score between RGB and gray-scale face images recovered from original events; ***Privacy*** measure face verification score between RGB and gray-scale images recovered from anonymized events. The area under the curve (AUC) scores in Figure 5 for ***RGB*** is 0.748, and ***No Privacy*** is 0.699 while for ***Privacy*** images it is 0.517. This suggests that the ArcFace model performs poorly on the images reconstructed from the anonymized event stream as the AUC=0.517, which is close to the random performance (AUC=0.50).

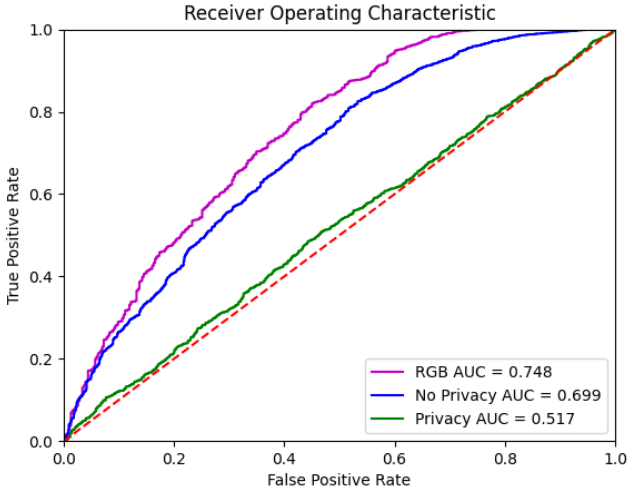


FIGURE 5. Face recognition accuracy using Arcface [65] model.

TABLE 2. Recovered image quality: SSIM and PSNR values.

Real Data	*SSIM*↓	*PSNR*↓
No-privacy	0.548	11.617
Privacy (Our)	0.384	8.943
Sythetic Data		
No-privacy	0.530	11.284
Privacy (Our)	0.368	8.071

Additionally, Table 2 presents image quality measurements using SSIM and PSNR. The results show that lower SSIM and PSNR scores are associated with degraded images. Considering the performance score of all three tests, image retrieval, face recognition, and image quality assessment, it is evident that our event anonymization network effectively preserves the anonymity of individual identity information.

Person ReId Performance: The rank accuracy and mean average precision score of person ReId utilizing anonymized event data for both real Event-ReId and SoftBio [66] datasets are presented in Figure 6. The results show that the anonymization model does not affect the person ReId performance. At the same time, shifting from ***No Privacy*** to ***Privacy-preserving*** the rank-1 accuracy and mAP drop is 5.5% and 5.2%, respectively for Event-ReId data. Similarly,

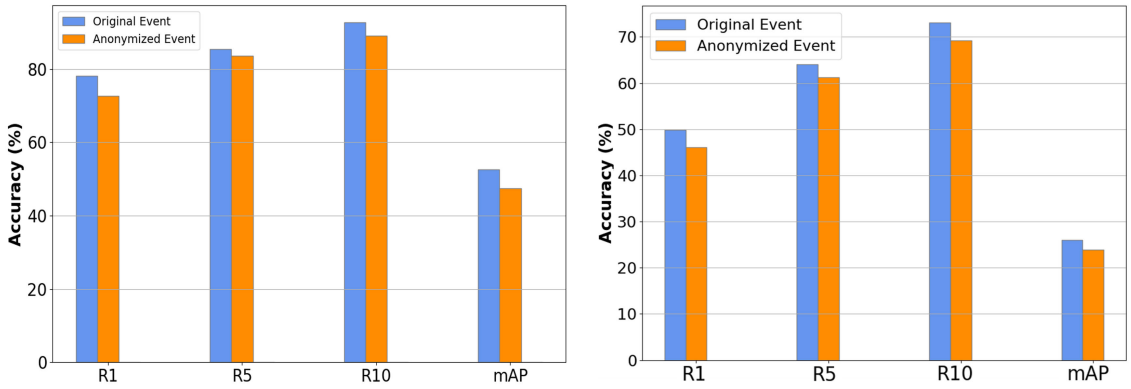


FIGURE 6. Person ReId performance using raw event blue and anonymized event orange, Event-ReId-v2 dataset (left), and SoftBio dataset (right).

TABLE 3. Person ReId performance between Event-ReId-v1 and Event-ReId-v2 dataset.

Dataset	R1	R5	R10	mAP
Event-ReId-v1 [15]	59.2	76.1	84.1	36.1
Event-ReId-v2	72.7	83.6	89.2	47.5

for SoftBio data, the drop in rank-1 is 3.8%, and for mAP is 2.1%.

Comparison With Event-ReId-v1: We also compared person ReId performance score with the previous version of the ReId dataset, **Event-ReId-v1** [15], which is a relatively small dataset. Table 3 shows an increase of 13.5% in rank-1 accuracy and 11.3% in mAP using the **Event-ReId-v2** dataset. This implies that providing more training data person ReId performance could further enhance.

Comparison With Baselines: Since this work investigates privacy-preserving person ReId for the first time, there is no other method for direct comparison. We benchmark our approach against event encryption (partial scrambling and partial discarding) methods [12] to check their effect on privacy-preserving. To perform event encryption, we use partial (75%) encryption for both the scrambling and discarding algorithms, as complete encryption distorts the entire visual information in the event data, which can not be utilized for downstream tasks. Table 4 reports SSIM and PSNR image quality metrics, $R1_{reid}$: Rank1 accuracy of person ReId, and $R1_{retr}$: Rank1 accuracy of image retrieval on reconstructed images, using proposed Event ReId dataset. Our proposed event anonymization method outperforms the event encryption technique.

Further, we compare the downstream task (person ReId) with a baseline Event-driven ReId method (Ed-ReId) [7]. Results in Table 5 illustrate that even after event-stream anonymization with our proposed network, person ReId performance is still better than Ed-ReId, although we pay a reasonable decrease in the score when applying privacy (i.e., the anonymization module).

Losses Ablation: We finally analyze the effect of losses on the downstream task accuracy (Person ReId). In the case of without privacy, if we remove the event anonymization module (α=β=0), the Rank1 accuracy is 78.2%, and with privacy, when we adopt the anonymization model with image reconstruction loss L_{rec} (α= 0, β=1) only, Rank1 accuracy significantly decreased to 69.1%. Finally, including L_{struct} loss (α=β=1) (which helps to maintain structural information while anonymizing the voxel-grid) recovers the accuracy to 72.7%, still preserving privacy, as detailed in Table 6 and Figure 7.

TABLE 4. Comparison with other methods on the Event-ReId dataset: image quality, event-reid, and image retrieval.

Method	$SSIM\downarrow$	$PSNR\downarrow$	$R1_{reid}\uparrow$	$R1_{retr}\downarrow$
No-privacy (raw events)	0.551	11.694	78.2	82.5
Encryption$_{Discarding}$ [12]	0.471	10.18	40.1	42.8
Encryption$_{Scrambling}$ [12]	0.436	10.06	31.0	32.6
Privacy (Ours)	**0.376**	**8.795**	**72.7**	**8.7**

TABLE 5. Comparison of person Re-Id performance on Event-ReId dataset with baseline.

Method	Privacy	R1	R5	R10	mAP
ED-ReId [7]	No	71.9	84.1	87.6	46.8
Ours	No	78.2	85.5	92.7	52.7
Ours	Yes	72.7	83.6	89.2	47.5

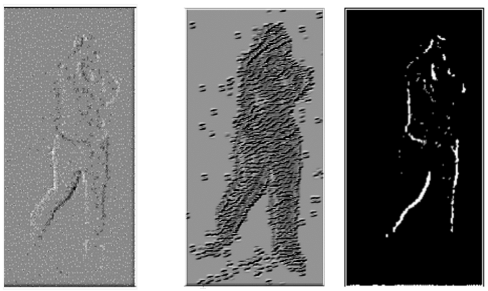


FIGURE 7. Raw event-voxel (left), anonymized event-voxel without L_{struct} loss (middle) and with L_{struct} loss (right).

TABLE 6. Ablation on the losses for person ReId accuracy.

$\alpha\mathcal{L}_{struct} + \beta\mathcal{L}_{rec} + \gamma\mathcal{L}_{reid}$	Method	R1	R5	R10
$\alpha = \beta = 0, \gamma = 1$	Raw_{Ev}	78.2	85.5	92.7
$\alpha = 0, \beta = \gamma = 1$	anonymized_{Ev}	69.1	81.8	87.3
$\alpha = \beta = \gamma = 1$	anonymized_{Ev}	72.7	83.6	89.2

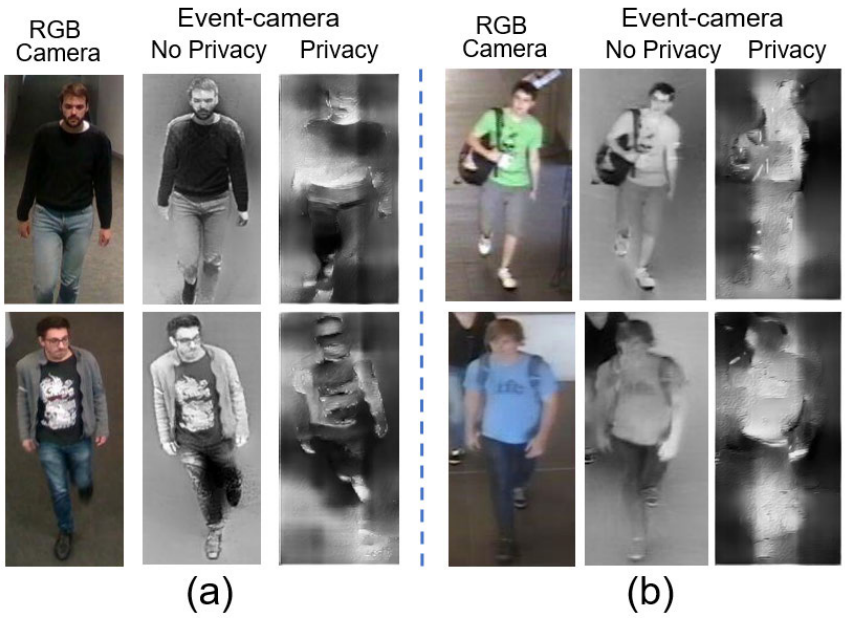


FIGURE 8. Visualization of reconstructed images obtained using raw and anonymized event data. a) real event dataset Event-ReId; b) simulated event dataset SoftBio.

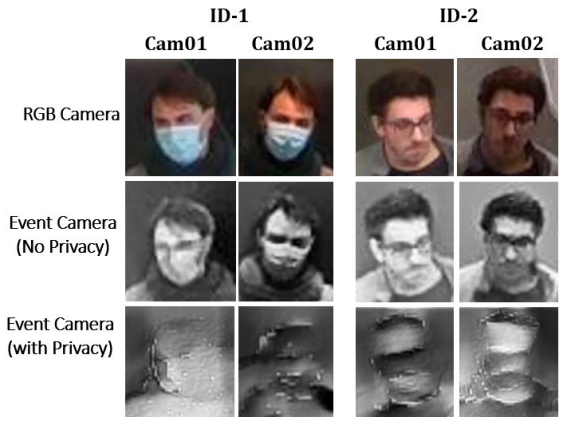


FIGURE 9. Visualization of reconstructed face images. Top row RGB original images. The middle row contains recovered images from the raw event. The bottom row contains recovered images from anonymized events.

Qualitative Results: We qualitatively compare the reconstructed images acquired using our approach with the original images. We show the results on two examples from each Event-ReId and Softbio data video from the dataset. Figure 8 displays anonymized images compared to the original RGB and recovered gray-scale images for reference. As observed, the image reconstructed from anonymized events degraded compared to non-privacy images. We also show the two exemplar face reconstructions from real event data in Figure 9, showing that the subject face can not be reconstructed from our anonymized event stream compared to face reconstruction from the non-privacy event stream.

D. HPE EXPERIMENTAL SETUP AND DATASET

Dataset: We also evaluate our proposed method on human pose estimation, using event-based human pose dataset DHP19 [2]. This is the first large-scale Human Pose Estimation (HPE) dataset captured from event cameras. The dataset contains 33 movement recordings of 17 subjects of different sex, age, and size. Each subject is recorded in a clutter-free indoor environment with four synchronized cameras having a resolution of 260 × 344 pixels, positioned at four different angles around the subject. The range of motions is narrow; most actions, such as leg kicking and arm abductions, are performed on the spot, except for slow running and walking. It includes 3D annotation as well as the camera parameters for the estimation of 2D projections.

HPE Evaluation Metric: We employed the mean per joint position error (MPJPE) metric for HPE evaluation. MPJPE is frequently used in HPE and is computed by calculating the average Euclidean distance between ground truth and predicted body joints. Moreover, MPJPE can be computed for both 2D and 3D HPE as:

$$\text{MPJPE} = \frac{1}{P}\sum_{n=1}^{P}\left\|\mathbf{h}_n - \hat{\mathbf{h}}_n\right\| \tag{13}$$

where P represents the number of the body joints, and h_n and $\hat{h}_n$ are the ground-truth and predicted (2D or 3D) position of n-th joint respectively.

Experimental Setup: Following [2], we use 12 subjects for training and 5 for testing. We set the size of the bin B=5; however, unlike the spatiotemporal voxel grid in person ReId, which is integrated using fixed time duration, we choose a fixed number of events (N = 7.5K) to generate a voxel grid because the average position of the 3D joint label of the corresponding event constant-count frame. The deep neural network input voxel grid size is 5 × 260 × 344. We set the batch size to 32 and trained the model with a learning rate of 0.0003 for 30 epochs.

FIGURE 10. Visualisation of reconstructed images obtained using raw event data (left) and anonymized event data (right) from the DHP19 dataset.

E. HPE RESULTS

Privacy-Preserving Performance: We also verify the proposed event anonymization technique effectiveness on the pose estimation dataset DHP19 [2] using an image retrieval test. As the DHP19 dataset is collected for pose estimation, we can still utilize it for image retrieval experiments because it is recorded with four event cameras placed at different angles around a subject performing various activities. Moreover, there is no corresponding ground truth RGB images available; DHP19 contains only event stream data. Therefore, we reconstructed images through E2VID from raw events to substitute RGB ground truth (see Figure 10.

We employ the same query-gallery setting of the Sec. C, "$Q_{no-privacy} \Leftrightarrow G_{privacy}$" that measures the image retrieval score on the anonymized (privacy-preserving) image gallery using recovered gray-scale query images from raw events. Similarly, for the comparison, the testing approach "$Q_{no-privacy} \Leftrightarrow G_{no-privacy}$" measures the image retrieval score on gallery images recovered from raw events. The retrieval accuracy is random for anonymized images (see Figure 11); this test also validates that our proposed method successfully preserves identity information.

Pose Estimation Performance: Table 7 illustrates the 2D test set results, defined as MPJPE (in pixel). Since event-based pose estimation methods [2] compute 2D prediction error only on front cameras (camera 2 and camera 3) of the DHP19 dataset. Thus, we compute the prediction error on cameras 2 and 3 for raw and anonymized voxel-grid. Table 7 shows the difference between the average error in the 2D joint position of raw and anonymized voxel is less than 1 pixel (came2: 0.32 and cam 3:0.27). Table 8 reports the 3D MPJPE score on test subjects. The difference in 3D prediction error is approximately 5mm between raw and anonymized voxel-grid. Both 2D and 3D HPE quantitative results and examples of qualitative results shown in Figure 12 and 13) suggest that the event stream can be anonymized without compromising the pose estimation accuracy to prevent privacy attacks (event-to-image-reconstruction).

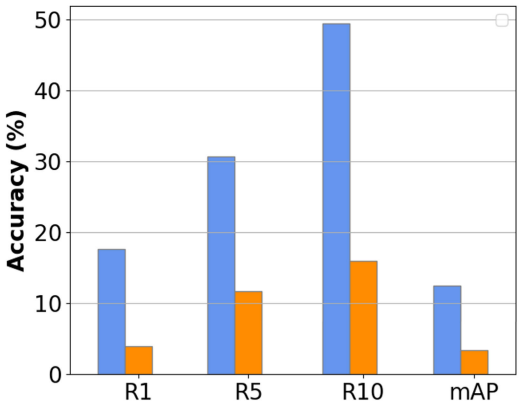


FIGURE 11. Image retrieval evaluation on DHP19 dataset, following query-gallery setting, blue: query[no privacy], gallery[no privacy], orange: query[no privacy], gallery[privacy].

TABLE 7. We report the 2D Mean Per Joint Precision Error (MPJPE, in pixels) of our proposed HPE method without privacy (raw event voxel) and with privacy (anonymized event voxel).

Camera View	No Privacy	Privacy
Camera 2	6.33	6.65
Camera 3	6.14	6.41

TABLE 8. We report the 3D Mean Per Joint Precision Error (MPJPE, in mm) of our proposed HPE (stereo) method without privacy (raw event voxel) and with privacy (anonymized event voxel).

Method	MPJPE(mm)
No Privacy	67.29
Privacy	72.53

TABLE 9. Comparison of 2D MPJPE (pixels) score with baseline methods.

Camera View	DHP19 [2]	Mnet [67]	**No Privacy (ours)**	**Privacy (ours)**
Camera 2	7.18	-	6.33	6.65
Camera 3	6.87	-	6.14	6.41
Mean	7.03	6.28	6.24	6.53

TABLE 10. Comparison of our 3D HPE method (stereo and monocular) with baseline methods using MPJPE (mm).

Method	input	MJJPE(mm)
DHP19 [2]	stereo	79.63
Privacy (Ours)	stereo	72.53
LiftMono [3]	monocular	95.51
Privacy (Ours)	monocular	101.47

Comparison With Baseline: We compare 2D and 3D pose estimation results obtained through our proposed network with baseline approaches. Table 9 and 10 illustrate a comparison of the 2D and 3D tests, respectively. For the 2D pose estimation score, we outperformed the DHP19 [2] method, while our score is comparable to the Mnet [67] approach. For the 3D HPE, we compute results for both *stereo* [2] and *monocular* [3] methods. Comparison in Table 10 shows our proposed method surpasses [2] in *stereo* settings. However, for *monocular* settings, our HPE performance score is lower than [3]; the difference is 5.96 MPJPE(mm).

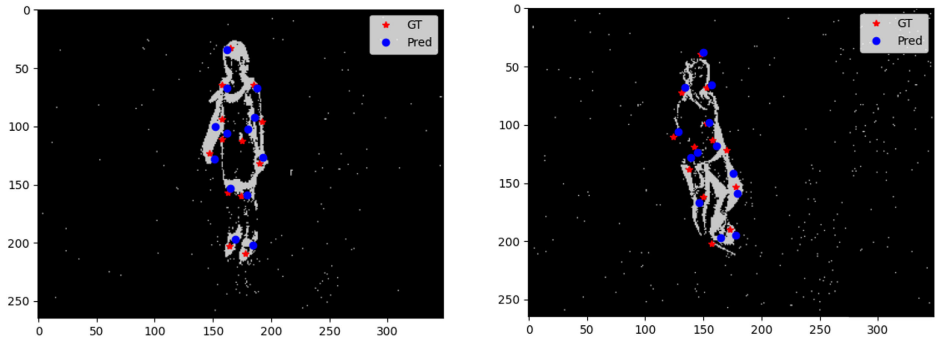


FIGURE 12. Visualization 2D human body joints estimation results, ground truth in red vs. prediction in blue: jumping (left), walking (right).

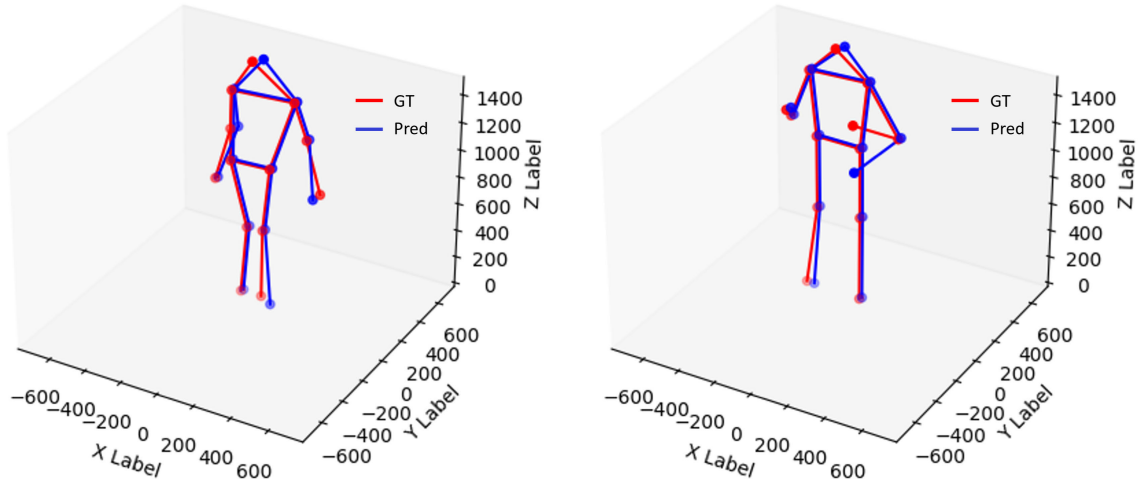


FIGURE 13. Visualization 3D human pose estimation results, ground truth pose in red vs. predicted pose in blue.

F. GENDER PREDICTION

The risk of extracting gender attributes, considered sensitive private information, poses another potential privacy attack. An attacker might aim to determine an individual's gender (male or female) from reconstructed gray-scale images. To assess the vulnerability of our system to such attacks, we analyze whether gender information can be recognized from the reconstructed images. We define this attack as a binary classification problem, where the task is for a gender classifier network to analyze the reconstructed image and predict the individual's gender.

For this purpose, we employ a state-of-the-art gender classification method known as MiVOLO [68]. This network is trained not only on facial features but also on broader aspects of a person's image, such as clothing and body shape, which could potentially offer additional clues about the individual's gender. By utilizing both these types of data, the classifier aims to provide a more comprehensive and accurate gender prediction.

We report the person detection and gender classification accuracy in Table 11 and 12 for real and synthetic event datasets, respectively. For the Event-ReId dataset, the MiVOLO gender classification method detected only 33.3% of persons and 26.7% identified gender accurately. The performance on the SOftbio dataset was even lower, with 18.4% detection and 6.6% gender classification accuracy. These outcomes indicate that our framework effectively conceals gender information in reconstructed images, thereby significantly enhancing privacy in event-based vision applications.

TABLE 11. Person detection and gender classification score using Event-ReId real data.

Method	*Detection Acc*	*Gender Acc*	*Male Acc*	*Female Acc*
RGB	100 (60/60)	93.3 (56/60)	96.1 (49/51)	77.8 (7/9)
No Privacy	100 (60/60)	78.3 (47/60)	88.2 (45/51)	22.2 (2/9)
Privacy	**33.3 (20/60)**	**26.7 (16/60)**	**29.4 (15/51)**	**11.1 (1/9)**

TABLE 12. Person detection and gender classification score using SoftBio synthetic dataset.

Method	*Detection Acc*	*Gender Acc*
RGB	88.8 (135/152)	54.0 (82/152)
No Privacy	81.0 (123/152)	52.0 (79/152)
Privacy	**18.4 (28/152)**	**06.6 (10/152)**

G. INVERSION ATTACK

We investigate a scenario where an attacker gains access to our privacy-preserving event camera and creates a large dataset comprising both anonymized event data and their

TABLE 13. SSIM, PSNR, and image retrieval score for recovered image under inversion attack and adversarial learning using Event-ReId dataset.

Method	*SSIM*↓	*PSNR*↓	$R1_{retr}$↓
No-Privacy	0.551	11.694	82.5
Privacy (Ours)	**0.376**	**8.795**	**8.7**
Inversion Attack	0.386	9.013	10.1
Adver. Learning	***0.419***	***9.482***	***15.4***

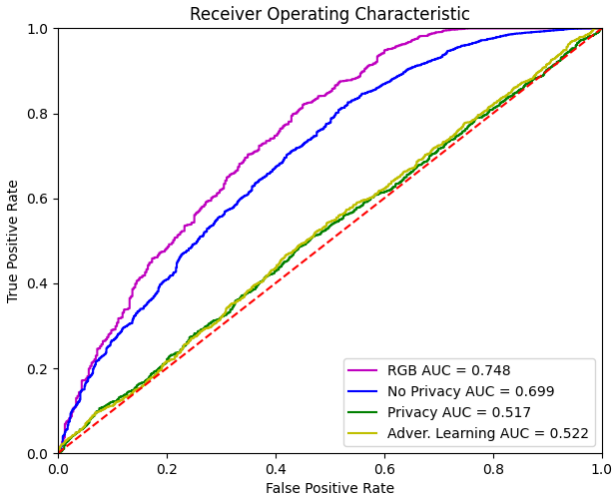


FIGURE 14. Face recognition accuracy for adversarial learning attack using Arcface [65] model.

original counterparts. In this situation, the attacker might train a network, denoted as E_{inv}, to undo the anonymization performed by E_{inv}, potentially resulting in the recovery of high-quality gray-scale images. We refer to this scenario as an ''inversion attack''. To assess how well our proposed framework withstands such privacy breaches, we train an autoencoder network using the real event dataset in a manner similar to the training of the E_{an} network. This network processes anonymized event streams output by the pre-trained E_{an} network and aims to minimize image reconstruction loss.

The quantitative results, shown in Table 13, measure the image retrieval and the image quality test (SSIM and PSNR) scores using these reconstructed images obtained through an inversion attack. The numbers indicate reconstruction quality is substantially poor, effectively preserving identity information. The qualitative results, shown in Figure 15(c) with two sample images, demonstrate that the reconstruction process fails to restore the images accurately. Consequently, the inversion attack is unable to counteract the anonymization effects of our event-based framework.

H. ADVERSARIAL LEARNING

We also consider a different scenario where an attacker has access not only to the anonymized events but also to the corresponding RGB/gray-scale images. In this situation, the attacker might attempt to retrain the image reconstruction module to produce high-quality images from the anonymized event data. This type of potential privacy attack is referred to as ''adversarial learning.'' To evaluate the resilience of

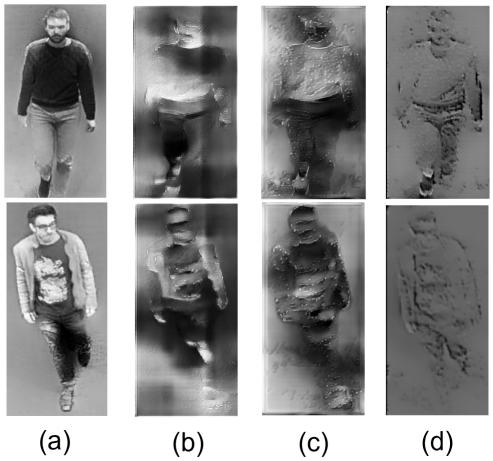


FIGURE 15. Reconstructed images from a) raw events, b) anonymized events, c) output of Inversion Attack, and d) output of Adversarial Learning.

our anonymization model against such a privacy attack, we retrained the image reconstruction model E_{rec} while keeping the rest of the pipeline fixed to determine if it can effectively extract details from the anonymized event voxels. For this retraining, we utilize the output anonymized event voxel from the pre-trained E_{an} model along with ground-truth data to retrain the E2VID network [8].

We evaluated the success of this retraining through various metrics. The SSIM and PSNR scores were used to assess the reconstructed image quality in Table 13. Besides, to evaluate privacy-preserving, we compute the image retrieval accuracy presented in Table 13 and the face recognition score in Figure 5. Furthermore, qualitative results shown in Figure 15(d) show the reconstruction efficacy. These evaluations collectively indicate that the reconstructed image quality remains significantly low, and the identity information is still preserved despite the adversarial learning attempt. That shows the framework effectively protects sensitive data against adversarial learning.

VI. CONCLUSION

In this work, we address the challenge of privacy in the context of neuromorphic vision sensors. Despite their inability to capture detailed RGB images, the potential of deep learning models to reconstruct high-quality images from event data poses a privacy threat. To cope with this, we have developed an end-to-end trainable network architecture that anonymizes event streams, ensuring privacy protection while maintaining the ability to perform downstream vision tasks. This network scrambles event data to degrade the quality of images that could be reconstructed, thus protecting privacy. We demonstrate how this framework can be set up for a person ReId and human pose estimation tasks, proving the framework's adaptability in various computer vision applications. We evaluate the proposed framework on both tasks with synthetic and real event-based datasets. We additionally propose and make available the first ever

person ReId neuromorphic dataset in order to motivate further progress in the field. Finally, our results on all datasets show that our approach effectively prevents possible privacy attacks on event data while executing person ReId and HPE tasks with a negligible performance impact. Collectively, this work marks a foundation for deploying ethical and privacy-aware surveillance technologies.

REFERENCES

[1] N. Messikommer, D. Gehrig, M. Gehrig, and D. Scaramuzza, "Bridging the gap between events and frames through unsupervised domain adaptation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3515–3522, Apr. 2022.

[2] E. Calabrese, G. Taverni, C. A. Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, and T. Delbruck, "DHP19: Dynamic vision sensor 3D human pose dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1695–1704.

[3] G. Scarpellini, P. Morerio, and A. Del Bue, "Lifting monocular events to 3D human poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1358–1368.

[4] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, "Event-based moving object detection and tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–9.

[5] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, "EKLT: Asynchronous photometric feature tracking using events and frames," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 601–618, Mar. 2020.

[6] G. Lenz, S.-H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Frontiers Neurosci.*, vol. 14, p. 587, Jul. 2020.

[7] S. Ahmad, G. Scarpellini, P. Morerio, and A. D. Bue, "Event-driven Re-ID: A new benchmark and method towards privacy-preserving person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 459–468.

[8] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1964–1980, Jun. 2021.

[9] F. Paredes-Vallés and G. C. H. E. de Croon, "Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3445–3454.

[10] W. Weng, Y. Zhang, and Z. Xiong, "Event-based video reconstruction using transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2543–2552.

[11] L. Zhu, X. Wang, Y. Chang, J. Li, T. Huang, and Y. Tian, "Event-based video reconstruction via potential-assisted spiking neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3584–3594.

[12] B. Du, W. Li, Z. Wang, M. Xu, T. Gao, J. Li, and H. Wen, "Event encryption for neuromorphic vision sensors: Framework, algorithm, and evaluation," *Sensors*, vol. 21, no. 13, p. 4320, Jun. 2021.

[13] T. Orekondy, B. Schiele, and M. Fritz, "Towards a visual privacy advisor: Understanding and predicting privacy risks in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3706–3715.

[14] S. Kumawat and H. Nagahara, "Privacy-preserving action recognition via motion difference quantization," 2022, *arXiv:2208.02459*.

[15] S. Ahmad, P. Morerio, and A. Del Bue, "Person re-identification without identification via event anonymization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11132–11141.

[16] B. R. Pradhan, Y. Bethi, S. Narayanan, A. Chakraborty, and C. S. Thakur, "N-HAR: A neuromorphic event-based human activity recognition system using memory surfaces," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.

[17] Q. Liu, D. Xing, H. Tang, D. Ma, and G. Pan, "Event-based action recognition using motion information and spiking neural networks," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1743–1749.

[18] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. D. Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner, and D. Modha, "A low power, fully event-based gesture recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7388–7397.

[19] L. Berlincioni, L. Cultrera, C. Albisani, L. Cresti, A. Leonardo, S. Picchioni, F. Becattini, and A. Del Bimbo, "Neuromorphic event-based facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4108–4118.

[20] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "HATS: Histograms of averaged time surfaces for robust event-based object classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1731–1740.

[21] J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "DDD17: End-To-End Davis driving dataset," 2017, *arXiv:1711.01458*.

[22] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. 2nd Conf. Robot Learn. (CoRL)*, 2018, pp. 969–982.

[23] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck, "DVS benchmark datasets for object tracking, action recognition, and object recognition," *Frontiers Neurosci.*, vol. 10, p. 405, Aug. 2016.

[24] G. Orchard, A. Jayawant, G. K. Cohen, and N. Thakor, "Converting static image datasets to spiking neuromorphic datasets using saccades," *Frontiers Neurosci.*, vol. 9, p. 437, Nov. 2015.

[25] 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," *Int. J. Robot. Res.*, vol. 36, no. 2, pp. 142–149, Feb. 2017.

[26] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3583–3592.

[27] C. Hinojosa, J. C. Niebles, and H. Arguello, "Learning privacy-preserving optics for human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2553–2562.

[28] C. Hinojosa, M. Marquez, H. Arguello, E. Adeli, L. Fei-Fei, and J. Carlos Niebles, "PrivHAR: Recognizing human actions from privacy-preserving lens," 2022, *arXiv:2206.03891*.

[29] Z. Wu, H. Wang, Z. Wang, H. Jin, and Z. Wang, "Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2126–2139, Apr. 2022.

[30] Z. Tasneem, G. Milione, Y. H. Tsai, X. Yu, A. Veeraraghavan, M. Chandraker, and F. Pittaluga, "Learning phase mask for privacy-preserving passive depth estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 504–521.

[31] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," *Comput. Secur.*, vol. 110, Nov. 2021, Art. no. 102402.

[32] X. Wang, H. Xue, X. Liu, and Q. Pei, "A privacy-preserving edge computation-based face verification system for user authentication," *IEEE Access*, vol. 7, pp. 14186–14197, 2019.

[33] R. S. Peres, A. Manta-Costa, and J. Barata, "Implementing privacy-preserving and collaborative industrial artificial intelligence," *IEEE Access*, vol. 11, pp. 74579–74589, 2023.

[34] Z. Wang, G. Yang, H. Dai, and C. Rong, "Privacy-preserving split learning for large-scaled vision pre-training," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1539–1553, 2023.

[35] F. Becattini, F. Palai, and A. D. Bimbo, "Understanding human reactions looking at facial microexpressions with an event camera," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9112–9121, Dec. 2022.

[36] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[37] G. Munda, C. Reinbacher, and T. Pock, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1381–1393, Dec. 2018.

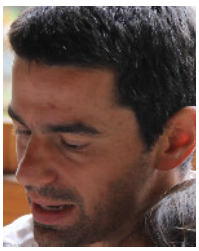
[38] Z. Yang, X. Jin, K. Zheng, and F. Zhao, "Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14278–14287.

[39] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[40] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 10254–10263.

[41] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 4610–4617.
[42] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 433–442.
[43] M. Paolanti, L. Romeo, D. Liciotti, A. Cenci, E. Frontoni, and P. Zingaretti, "Person re-identification with RGB-D camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection," *Sensors*, vol. 18, no. 10, p. 3471, Oct. 2018.
[44] S. Dou, X. Jiang, Q. Zhao, D. Li, and C. Zhao, "Towards privacy-preserving person re-identification via person identify shift," 2022, *arXiv:2207.07311*.
[45] J. Dietlmeier, J. Antony, K. McGuinness, and N. E. O'Connor, "How important are faces for person re-identification?" in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6912–6919.
[46] B. Zhao, Y. Li, X. Liu, H. H. Pang, and R. H. Deng, "FREED: An efficient privacy-preserving solution for person re-IDentification," in *Proc. IEEE Conf. Dependable Secur. Comput. (DSC)*, Jun. 2022, pp. 1–8.
[47] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation," *IEEE Access*, vol. 8, pp. 133330–133348, 2020.
[48] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, "2D human pose estimation: A survey," *Multimedia Syst.*, vol. 29, no. 5, pp. 3115–3138, Oct. 2023.
[49] J. Stenum, K. M. Cherry-Allen, C. O. Pyles, R. D. Reetzke, M. F. Vignos, and R. T. Roemmich, "Applications of pose estimation in human health and performance across the lifespan," *Sensors*, vol. 21, no. 21, p. 7315, Nov. 2021.
[50] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[51] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, Dec. 2019.
[52] G. Lan, Y. Wu, F. Hu, and Q. Hao, "Vision-based human pose estimation via deep learning: A survey," *IEEE Trans. Hum.-Mach. Syst.*, vol. 53, no. 1, pp. 253–268, Feb. 2023.
[53] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, and C. Theobalt, "EventCap: Monocular 3D capture of high-speed human motions using an event camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4967–4977.
[54] R. W. Baldwin, R. Liu, M. Almatrafi, V. Asari, and K. Hirakawa, "Time-ordered recent event (TORE) volumes for event cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2519–2532, Feb. 2023.
[55] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.
[56] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5419–5427.
[57] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 989–997.
[58] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4277–4281.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[60] H. Qu, L. Xu, Y. Cai, L. G. Foo, and J. Liu, "Heatmap distribution matching for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24327–24339.
[61] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "3D human pose estimation with 2D marginal heatmaps," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1477–1485.
[62] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.
[63] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.
[64] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.
[65] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.
[66] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey, "A database for person re-identification in multi-camera surveillance networks," in *Proc. Int. Conf. Digit. Image Comput. Techn. Appl. (DICTA)*, Dec. 2012, pp. 1–8.
[67] G. Goyal, F. Di Pietro, N. Carissimi, A. Glover, and C. Bartolozzi, "MoveEnet: Online high-frequency human pose estimation with an event camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4023–4032.
[68] M. Kuprashevich and I. Tolstykh, "MiVOLO: Multi-input transformer for age and gender estimation," 2023, *arXiv:2307.04616*.

SHAFIQ AHMAD received the B.Sc. degree in electronics engineering from COMSATS University Islamabad, Abbottabad Campus, Pakistan, in 2011, and the M.S. degree in electrical engineering from the Institute of Space Technology, Islamabad, Pakistan, in 2016. He is currently pursuing the Ph.D. degree with the Pattern Analysis and Computer Vision (PAVIS) Laboratory, Istituto Italiano di Tecnologia at Italy (IIT), collaborating with the Department of Naval, Electrical, Electronic, and Telecommunications Engineering, University of Genoa. His current research interest includes deep learning for event-based privacy-preserving visual surveillance (person re-identification and human pose estimation).



PIETRO MORERIO (Member, IEEE) received the B.Sc. and M.Sc. degrees (summa cum laude) in physics from the University of Milan, Italy, in 2007 and 2010, respectively, and the Ph.D. degree in computational intelligence from the University of Genoa, Italy. He was a Research Fellow in video analysis for interactive cognitive environments with the University of Genoa, from 2011 to 2012. From 2016 to 2021, he was a Postdoctoral Researcher with Istituto Italiano di Tecnologia at Italy (IIT), Genoa, Italy, where he is currently a Technologist with the Pattern Analysis and Computer Vision (PAVIS) Research Line. His research interests include machine learning, deep learning, and computer vision.

ALESSIO DEL BUE (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, Queen Mary University of London. He is currently a Tenured Senior Researcher leading with the Pattern Analysis and Computer Vision Research Line, Istituto Italiano di Tecnologia at Italy (IIT), Genoa, Italy. Previously, he was a Researcher with the Institute for Systems and Robotics, Instituto Superior Técnico (IST), Lisbon, Portugal. He is the coauthor of more than 100 scientific publications in refereed journals and international conferences. His current research interest includes 3D scene understanding from multi-modal input (images, depth, and audio) to support the development of assistive AI systems. He is an ELLIS Member of the Genoa Unit and a member of the technical committees of important computer vision conferences (CVPR, ICCV, ECCV, and BMVC). He serves as an Associate Editor for *Pattern Recognition* and *Computer Vision and Image Understanding* journals.

Open Access funding provided by 'Istituto Italiano di Tecnologia' within the CRUI CARE Agreement