Neurocomputing 547 (2023) 126388

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Neuromorphic high-frequency 3D dancing pose estimation in dynamic environment

Zhongyang Zhang [a], Kaidong Chai [b], Haowen Yu [b], Ramzi Majaj [b], Francesca Walsh [b], Edward Wang [a], Upal Mahbub [c], Hava Siegelmann [b], Donghyun Kim [b], Tauhidur Rahman [a,*]

[a] University of California San Diego, 9500 Gilman Dr, La Jolla, 92093, CA, USA
[b] University of Massachusetts Amherst, 181 Presidents Drive, Amherst 01003, MA, USA
[c] Qualcomm Technologies, Inc., San Diego, CA, USA

ARTICLE INFO

Article history:
Received 26 January 2023
Accepted 22 May 2023
Available online 26 May 2023
Communicated by Zidong Wang

Keywords:
Event Camera
Dynamic Vision Sensor
Neuromorphic Camera
Simulator
Dataset
Deep Learning
Human Pose Estimation
3D Human Pose Estimation
Technology-Mediated Dancing

ABSTRACT

Technology-mediated dance experiences, as a medium of entertainment, are a key element in both traditional and virtual reality-based gaming platforms. These platforms predominantly depend on unobtrusive and continuous human pose estimation as a means of capturing input. Current solutions primarily employ RGB or RGB-Depth cameras for dance gaming applications; however, the former is hindered by low-light conditions due to motion blur and reduced sensitivity, while the latter exhibits excessive power consumption, diminished frame rates, and restricted operational distance. Boasting ultra-low latency, energy efficiency, and a wide dynamic range, neuromorphic cameras present a viable solution to surmount these limitations. Here, we introduce *YeLan*, a neuromorphic camera-driven, three-dimensional, high-frequency human pose estimation (HPE) system capable of withstanding low-light environments and dynamic backgrounds. We have compiled the first-ever neuromorphic camera dance HPE dataset and devised a fully adaptable motion-to-event, physics-conscious simulator. YeLan surpasses baseline models under strenuous conditions and exhibits resilience against varying clothing types, background motion, viewing angles, occlusions, and lighting fluctuations.

1. Introduction

Technology-Mediated Dancing (TMD), which leverages digital systems to facilitate remote, engaging, and health-promoting physical activities in dance form, has always been a part of gaming. More and more so, TMD is becoming a central part of the immersive experience as the digital experience becomes more and more blended with the physical reality. [1–3]. Ranging from remote control-based gaming console games to Virtual Reality (VR) platforms, various TMD forms are intertwining with users' living spaces. Human pose estimation (HPE) serves a crucial function in TMDs, as it discerns users' unique, intricate, and rapidly evolving dance poses, enabling interaction with computers. To accommodate the broadest range of players, TMD necessitates high-fidelity HPE that functions reliably in diverse, challenging, and realistic indoor environments, such as dynamic lighting and background conditions.[1]

Contemporary HPE systems predominantly rely on depth and RGB cameras; however, both standard RGB camera-based and depth-based monocular HPE systems [4–6] struggle to generate ultra-fast, high-speed human pose inferences due to their limited frame rates–a critical requirement for numerous real-world applications (e.g., virtual reality dance games, high-frequency motion characterization for tremor monitoring applications). The FPS of mainstream RGBD cameras (e.g., Microsoft Azure Kinect and Intel RealSense) is typically limited to 30 FPS due to a combination of hardware and software constraints, such as the time overhead introduced by IR projectors. Although some high-spec RGB cameras support relatively high FPS, they are usually fixed and have an attainable upper limit. In contrast, neuromorphic cameras can achieve a bandwidth of over 10 million events per second with a minimum latency of approximately 20 μs (DAVIS 346). Provided that computational resources are available, neuromorphic cameras

Abbreviations: DVS, Dynamic Vision Sensor; HPE, Human Pose Estimation; TMD, Technology-Mediated Dancing.

* Corresponding author.

[1] Datasets were solely downloaded and evaluated by UC San Diego.

https://doi.org/10.1016/j.neucom.2023.126388
0925-2312/

can attain extremely high-frequency inference. A detailed analysis can be found in the Results section surrounding Fig. 16. Moreover, while RGB-based HPE falters in low-light conditions due to substantial motion blur-related issues, depth cameras can only operate within a limited depth range. Furthermore, neither camera inherently distinguishes between static and moving objects, leading to both static and dynamic objects in the background being captured alongside the target human body. These shortcomings compromise HPE robustness in dynamic settings. Depth cameras, which employ active infrared light, also demand considerable power consumption. Consequently, our proposed work aims to develop a high-speed, low-latency, and low-power 3D human pose estimation framework capable of functioning in dynamic lighting and motion environments.

Neuromorphic cameras [7], also known as Dynamic Vision Sensors (DVS) or event cameras, have a silicon retina design based on a core mechanism of mammalian vision, rendering them particularly sensitive to moving targets and fluctuating lighting conditions. These cameras operate on the premise that, for mammals, moving objects often contain more valuable information for hunting and evading predators, whereas relatively static backgrounds deserve less attention for continuous monitoring and processing. Emulating this dynamic vision characteristic, each neuromorphic camera pixel operates asynchronously; independently monitoring logarithmic minute brightness changes, thereby ensuring sensitivity to motion in both high and low-light conditions. This mechanism inherently filters out static backgrounds without necessitating the transmission of detailed, bulky frames each time an event occurs. Instead, moving targets can rapidly activate numerous events with exceptional time resolution, enhancing neuromorphic camera sensitivity to dynamic targets. The wide dynamic range renders neuromorphic cameras resilient across various situations: from nighttime to glaring noon, from tunnels to night driving. Beyond their advantages in low-light conditions, neuromorphic cameras also exhibit reduced sensitivity to skin color and brightness changes [8,9].

DVS HPE has garnered considerable interest recently due to the aforementioned advantages, prompting the collection of several noteworthy datasets [10,11]. Despite these datasets utilizing fixed patterns for motion guidance rather than dances, they also exhibit significant limitations concerning the development of pragmatic systems. Data acquisition in these studies is conducted in ideal laboratory settings, devoid of any background motion to serve as interference. Lighting conditions are optimal and stable, precluding the exploration of low and/or variable lighting scenarios. These uncertainties and noise, reflective of real-world environments, present genuine challenges that must be addressed to render this technology more practical and generalizable. Consequently, we introduce two novel datasets: one featuring a real-world dynamic background under both high and low-light conditions, and the other comprising a simulated environment with a fully controllable and customizable pipeline for generating new data samples on demand. These datasets not only suit the needs of this research but also benefit the broader community as they will be made publicly available. A comparison between the existing neuromorphic camera-based HPE datasets and our assembled datasets are presented in Table 1.

Not only are existing datasets overly idealistic, previous DVS HPE efforts are further constrained by the "missing torso" problem, an intrinsic limitation stemming from the dynamic properties of the neuromorphic camera. When certain human body parts remain stationary, neuromorphic cameras solely capture other moving components, disregarding these static portions. Consequently, during such periods, the event representation contains minimal or no information about these parts, leading to substantial estimation inaccuracies.

To construct a DVS HPE system compatible with more realistic environments, we propose a two-stage system, *YeLan*[2], which accurately estimates human poses under low-light conditions amid noisy background elements. In the first stage, an early-exit-style mask prediction network is implemented to eliminate moving background objects while maximizing energy efficiency. The second stage employs a BiConvLSTM to facilitate information flow between frames, mitigating the missing torso issue. Additionally, TORE volume is utilized to construct denser and more informative input tensors, addressing the low event rate problem in low-light settings. We conducted extensive experiments comparing our approach with baseline DVS HPE methods, achieving state-of-the-art results on the two proposed new datasets.

In summary, the core contributions of this paper are as follows:

1. YeLan represents the first neuromorphic camera-based 3D human pose estimation solution tailored for dance motion, functioning robustly in challenging conditions such as low lighting and occlusion, and overcoming neuromorphic camera limitations while harnessing their strengths.
2. Our end-to-end simulator enables precise, low-level control over generated events and produces the first and largest neuromorphic camera dataset for dance HPE. This synthetic dataset (*Yelan-Syn-Dataset*) surpasses existing resources in quantity and variability.
3. We carry out a human subject study and collect a real-world dance HPE dataset (*Yelan-Real-Dataset*), taking into account low-light conditions and mobile background content. Both datasets and code will be made publicly available following publication.

2. Related Work

Dance and Its Effects: Spanning various age groups, dance has a rich history and is a beloved form of communication and physical activity with a rich history. Due to its cultural diversity and practicality, dance has evolved into numerous variations across regions and time periods, encompassing diverse styles, rhythms, intensities, and steps. Substantial research evidence supports the notion that dance exerts a significant positive influence on physical and mental health [15]. A recent study [16] revealed that dance can enhance neuroplasticity and stimulate neural activation in multiple brain regions. Consequently, dance can serve as a rehabilitative tool for an array of brain-related pathologies [15]. Recent literature indicates that dance therapy can have a considerable positive impact on depression [17], schizophrenia [2], Parkinson's [18], fibromyalgia [1,19], dementia [20], cognitive deterioration [3,21], stress [22], and chronic stroke [23].

Technology-mediated dance has gained popularity in recent years [24], rendering dance more accessible and facilitating learning [25,26]. Dance has been integrated into various video games [24,27,25,28], including the *Just Dance Series* [29], *Dance Dance Revolution* [30], and *Dance Central* [31]. Additionally, movement-based VR rhythm games such as *Beat Saber* [32], *Synth Riders* [33], and *Dance Collider* [34] have emerged.

High-fidelity 3D human pose estimation is a critical component of technology-mediated dance, as it enables the translation of dance gestures into input or commands [35,36] for AR/VR or dance video games. These 3D human pose estimators facilitate performance evaluation, personalized feedback, and choreography [24]. However, conventional RGB and depth camera-based 3D human pose estimations often struggle to capture rapid or high-speed

[2] Derived from a character in the video game Genshin Impact, the name *YeLan* signifies "night orchid" and can also be construed as the Chinese term for "night viewing."

Table 1
A comparison between existing 3D HPE datasets and our collected datasets. †: "Modality" indicates the modality used in the proposed pipeline rather than all paired modalities. ‡: "tight" refers to Mocap-specific tight clothing, "casual" to everyday attire, and "arbitrary" to any clothing type, even if it may cause significant occlusion. ∗: This dataset is not publicly available.

Study	Human3.6 M[12]	MKV[13]	DHP19[10]	MMHPSD∗[14]	YeLan	
Modality†	RGB	RGBD	DVS	DVS	DVS	
Inference Rate	50 FPS	10 FPS	Arbitrary, 50 FPS verified.	15FPS	Arbitrary, 150 FPS verified.	
Lighting	Ideal	Ideal	Ideal	Ideal	Low to High	
Data Type	Real	Real	Real	Real	Synthetic	Real
Background	Static	Static	Static	Static	Dynamic	Both
View	Arbitrary	Arbitrary	Arbitrary	Arbitrary	Arbitrary	Front
Clothing‡	Casual	Casual	Tight	Casual	Arbitrary	Tight
Data Size	3,600,000	22,406	350,860	240,000	3,958,169	446,158

pose changes during dance performances in challenging conditions (e.g., low light, dynamic background). In this work, we introduce a neuromorphic camera-based 3D human pose estimation system designed to support technology-mediated dance in these dynamic and demanding settings.

Human Pose Estimation: Depending on the taxonomy, existing 3D human pose estimation methods can be classified by modality, the number of sensors used in tracking, and result forms. Most approaches rely on RGB-based [37–39,6,5] or RGB-Depth-based modalities [40–42,13]. Generally, RGB image-based methods have lower equipment requirements and are more extensively explored. However, RGB-Depth cameras benefit from the additional depth information they provide, aiding in detection, segmentation, and parts localization. Still, depth cameras depend heavily on IR projectors for depth map construction, which consume significant power and are sensitive to bright environmental light. This limits the working distance and FOV of RGBD cameras and hinders their outdoor application.

For the second classification method, the resulting two categories are: monocular [43,5,13] and multi-view [44–46] methods. The multi-view method requires substantially more power and involves capturing subjects from multiple views/cameras positioned at different angles around a capturing area [47]. Moreover, setting up a multi-view neuromorphic camera data collection system is complex and costly, making applications outside the lab very challenging. The monocular method estimates human pose from a single camera view.

An alternative classification approach for human pose estimation techniques involves determining if they are model-based methods, as described by Omran et al. [48], or skeleton-based methods, as outlined by Chen et al. [4]. While the model-based method seeks to reconstruct the full 3D body shape of a human model [6], skeleton-based methods utilize a bone skeleton as an intermediate representation and regress the joint locations in 3D space. In this work, we use a monocular skeleton-based DVS HPE approach to achieve a more efficient and practical solution for 3D HPE in real-world settings.

Dynamic Vision Sensor (DVS), or neuromorphic camera, was originally proposed by Lichtsteiner et al. [7]. In recent years, the neuromorphic camera has increasingly garnered attention and has been applied to various computer vision tasks, including object recognition [49,50], segmentation [51], corner detection [52,53], gesture recognition [10,54,55], optical flow estimation [56,57], depth estimation [58,59], Simultaneous Localization And Mapping (SLAM) [60,61], autonomous driving [62,63], and human pose estimation [64,10,14]. While RGB cameras struggle due to motion blur, neuromorphic cameras are highly sensitive to lux variations in both extremely overexposed and underexposed scenes [65] by design.

Neuromorphic Camera-based Human Pose Estimation: While neuromorphic cameras have been previously proposed for human pose estimation in existing literature, the focus of prior work has mainly been on designing algorithms for relatively noise-free, background-activity-less, and well-lit settings. For instance, the recent neuromorphic camera HPE dataset, DHP19 [10], features event recordings of human movements, poses, and moving objects, and [14] further collected a multi-modality HPE dataset. Additionally, some neuromorphic camera datasets have been employed in gesture recognition [66] and action recognition [67]. However, these datasets capture a limited number of motion trajectory types in controlled, noise-free environments, which hampers the development of a robust neuromorphic camera-based sensing system.

The lack of a diversified dataset in realistic conditions under low and dynamic light is a major bottleneck for creating a more robust neuromorphic camera-based sensing system. In this work, we address this issue by demonstrating how a neuromorphic camera-based mobile sensing platform can effectively capture high-frequency 3D human poses during a dance performance while being highly resilient to various real-world challenging conditions. We aim to develop a versatile and robust system for human pose estimation that can be successfully applied in a wide range of real-world scenarios by incorporating resilience to low light; dynamic moving backgrounds; higher sensor fields of view; longer distances between the sensor and target human; and diverse outfits or clothing.

3. Design Consideration

Challenges in Neuromorphic Camera-based Dancing HPE: In contrast to everyday movements or gestures such as walking, jumping, and waving hands, dance encompasses a far greater diversity of movements that are unlikely to be included in existing fixed-activities-based HPE datasets, rendering dancing HPE a more challenging and generalized problem. Furthermore, compared to gesture recognition-level daily actions, a dancer constantly alters their pose at a significantly faster pace. The average joint location change rate is generally much larger, imposing a higher demand on the temporal resolution of pose estimation. This is particularly true for high-fidelity and professional dancing recordings employed in the gaming industry or virtual dance video production.

Neuromorphic cameras, with their inherent high-frequency and motion-sensitive characteristics, are well-suited for addressing these challenges. However, they also face unique obstacles compared to RGB-based methods. For instance, in noisy and dynamic environments, their extreme motion sensitivity can result in numerous events from environmental signals rather than the actual dancer's motion. Sometimes, these events even overwhelm those generated by the dancer, leading to a severe event filtering problem. Moreover, low-light conditions produce fewer events for the same movement and hence further lower the signal-to-noise ratio in the event representation.

To address these challenges, YeLan proposes an early-exit-style event filtering mechanism that predicts a binary mask over the target human body and rejects all events not generated due to the user's body movement. This helps to improve the accuracy and reliability of the human pose estimation model in various environments.

Missing torso problem: Another unique problem for DVS HPE is the missing torso problem. During physical activity, not all body parts move equally. For instance, the upper body may be in motion while the lower body remains stationary. This partial immobility results in the silence of corresponding pixels within the neuromorphic camera, leading to a higher error rate in predicting these missing joints. To solve this problem, we need to focus on the following points:

1. *Appropriate spatiotemporal representation:* An adeptly designed spatiotemporal representation can efficiently compile information derived from an extensive event stream, preserve vital historical data, and eliminate noise events. These factors collectively enhance the understanding of human body part locations and mitigate the missing torso issue.
2. *Information flow between frames:* Neighboring frames exhibit similarities in historical information and human joint locations. By considering the feature maps of these frames as a time series and employing a robust temporal method to facilitate information flow between frames, it becomes possible to gather insights regarding the missing torso and generate more persuasive joint location estimates.

Prior research on neuromorphic cameras has introduced various representations, such as event frames [68], constant-count frames [69], time surfaces [70], and voxel grids [71]. A time surface is a 2D representation in which each pixel encodes the timestamp of the most recent event as its pixel value. Each pixel retains the timestamp of the latest event [72] occurring at that location. A voxel grid, meanwhile, constitutes a 3D histogram of events, where each voxel denotes the number of events within a specified interval at a particular pixel location. By retaining the spatiotemporal information of the entire event history, voxel grids circumvent the loss of information associated with collapsing the history into a 2D grid representation. However, event frames, event counts, and voxel grids do not effectively conserve distantly related historical information, deteriorating the missing torso problem. Time surfaces, conversely, may discard temporal information and are unable to maintain information from multiple events at identical pixel locations. Furthermore, these existing representations are hindered by a low signal-to-noise ratio (SNR) in poorly illuminated conditions.

In this study, we employ a modified Time Ordered Recent Event (TORE) volume representation [73] that concurrently preserves both the most recent and relevant historical information, thereby helping alleviate the missing torso issue. This representation also functions as a noise filter for noise types such as salt and pepper without impacting other significant signals.

Spatiotemporal representation: Learning a representation that preserves meaningful spatiotemporal information properly about the human pose is a fundamental challenge. Especially given that neuromorphic cameras can generate a massive event sequence within a short period of time.

Prior work on neuromorphic cameras proposed different representations, including event frame [68], constant-count frame [69], time surface [70], and voxel grid [71]. Time surface is a 2D representation where each pixel records the most recent event's timestamp as its pixel value. Each pixel stores the timestamp of the last event [72] in that location. A voxel grid is a 3D histogram of events. Each voxel represents the number of events in an interval at a pixel location. Voxel grid prevents information loss by preserving spatial–temporal information of the whole history instead of collapsing the history into a 2D grid representation. However, event frame, event count, or voxel grid do not preserve relatively distant historical information, which gives rise to the missing torso problem. Time surface, on the other hand, can discard the temporal information and cannot keep the information from multiple events at the same pixel location. Moreover, these existing representations suffer from a low signal-to-noise ratio (SNR) in low-lighting conditions.

In this work, we use a modified Time Ordered Recent Event (TORE) volume representation [73] that can simultaneously preserve both the latest information and recent historical information, which can then compensate for the missing torso problem. It also serves as a noise filter for noises like salt and paper without affecting other informative signals.

4. Proposed System

4.1. Event Preprocessing and TORE Volume

TORE [73] attempts to mimic the human retina by preserving the membrane's potential properties. A fixed-length(e.g., K) First-In-First-Out (FIFO) queue is adopted to record the relative timestamp of the most recent K events. When a new event enters a pixel's queue, its relative timestamp is inserted, and the oldest event in the queue is expelled. TORE calculates the logarithm of these timestamps in the FIFO buffer. TORE transforms the sparse event stream into a dense, bio-inspired representation with minimal information loss, achieving state-of-the-art results in various DVS tasks (e.g., classification, denoising, human pose estimation). A comprehensive comparison of different representations can be found in [73].

In contrast to the original TORE volume, we modify TORE through normalization, 0–1 flip, and range scaling. We first perform the missing normalization and then invert the normalized maximum and minimum values. In the original paper [73], the oldest events and pixels with no recorded events are assigned the maximum value $log(\tau)$, which is counterintuitive and may hinder the convergence speed. We can readily address this issue by setting the value v in TORE to $1 - v/log(\tau)$. With this modification, older events have lower values, while newer events have higher values, and pixels with no recorded events are set to 0. Lastly, since the logarithm is utilized in TORE calculations, the most recent $4.63\mu s$ occupies the $[0.7, 1]$ range of the entire $[0, 1]$, and $4.63\mu s$ is shorter than the DVS camera's sensitivity time (150 ms, as used in the original TORE paper [73]). Consequently, a rescaling is introduced to mitigate this issue. Our modification formula is as follows:

$$v\prime = (1 - v/log(\tau))|_{[0,0.7]}/0.7 \tag{1}$$

where v and $v\prime$ denote the original and modified TORE values, while the notation $x|_{[a,b]}$ indicates the value of x is clamped within the range $[a, b]$.

TORE maintains a first-in-first-out queue individually for each pixel, with values in the queue decreasing over time. TORE can preserve pixel history for up to five seconds. These attributes render it suitable for dynamic conditions. In high-lighting conditions, TORE can discard redundant histories for the same pixel, alleviating the model's computational burden. In low-lighting conditions, past event histories can compensate and aid in determining the position of some joints that have not moved significantly during the last time window, thereby helping address the missing torso issue. Fig. 1 illustrates the process of generating modified TORE volumes. (See Fig. 2).

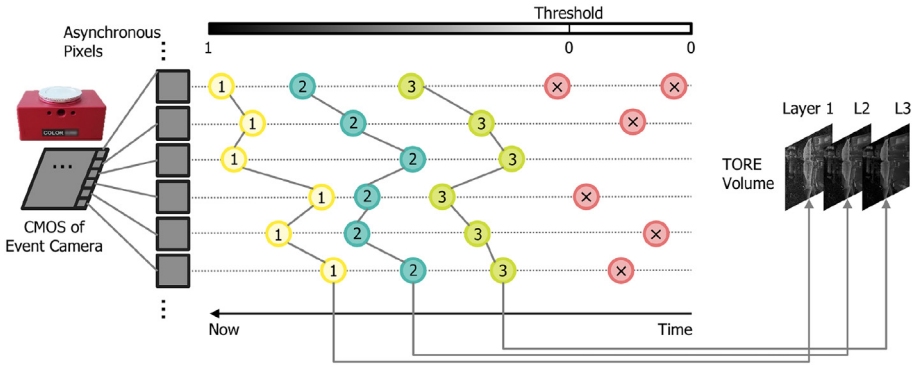

Fig. 1. The schematic representation of the TORE volume. As each pixel in the neuromorphic camera operates asynchronously, we establish a First-In-First-Out (FIFO) queue for each pixel corresponding to each event polarity (polarity denotes whether the event is triggered by an increase or decrease in brightness). The queue possesses a depth of K (here K = 3), and when a queue becomes full and subsequent events arrive, the oldest event is removed from the queue. The i^{th} layer of TORE is produced by extracting the i^{th} event stored in the FIFO queues of all pixels.

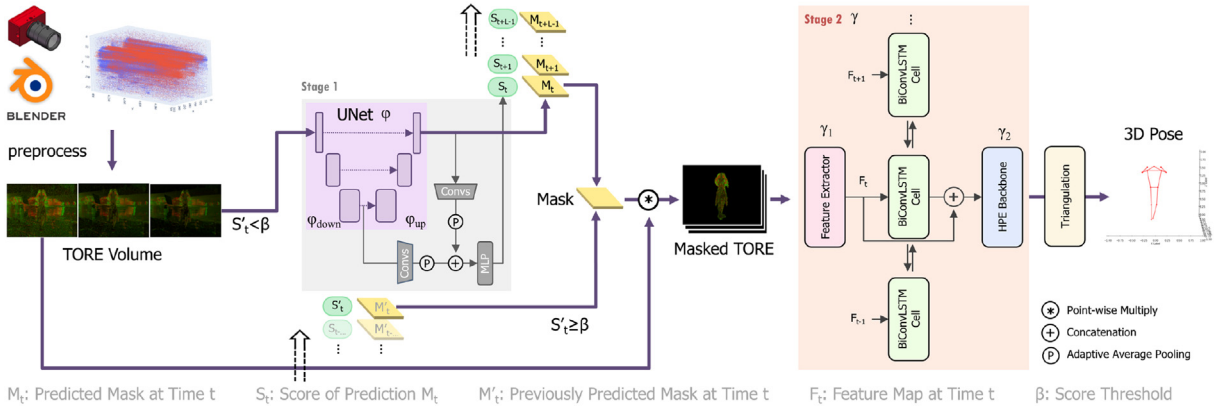

Fig. 2. The pipeline of the proposed *YeLan*. It initially processes the event stream into TORE volumes, which are subsequently sent to the stage one human body mask prediction network. This network predicts a series of masks for the ensuing frames, accompanied by quality-assessment scores to minimize computation costs. The estimated human mask undergoes point-wise multiplication with the original TORE volume before advancing to the next stage. Stage two encompasses the human pose estimation network, where BiConvLSTM and three hourglass-like refinement blocks are employed to estimate the heatmap of joints' projections on three orthogonal planes. The precise 3D coordinates of these joints are determined through a triangulation method based on these heatmaps.

4.2. Event Filtering with Human Body Mask Prediction Network

To filter out events triggered by background activities and neuromorphic camera hardware noise (e.g., hot pixel and leak noise [74]), a mask prediction network is employed, capable of predicting a human body mask from the TORE volume representation.

For the mask prediction network, we adopt a modified version of U-Net [75] proposed by Olaf et al. Unlike the original version, our modified iteration can predict the mask for the current input frame and a series of masks following this frame. Each predicted mask is generated in conjunction with a confidence score. The primary rationale behind predicting the human body mask of future frames is that future motion trajectories of different human body parts are generally predictable with information about current and previous motion trajectories. Time-Ordered Recent Event (TORE) volume representation efficiently captures current and previous motion history. This enables the mask prediction network to estimate human body masks of the current and future frames and their corresponding confidence scores. The confidence scores offer an early-exit-like mechanism, allowing the computational pipeline to bypass mask prediction for a specific frame if the mask predicted based on a previous frame's TORE volume achieves a high confidence score. This mechanism significantly reduces the computational cost and energy consumption (further results in Section 7.3). For the predicted mask, the U-Net generates floating point numbers between 0 and 1, with a binarization process employing a minimal threshold of 0.1. The small threshold ensures that the generated mask does not exclude any parts of the human body, as failure to encompass the entirety would result in greater error than permitting a slight amount of noise.

Denote the input TORE representation as X_t, and the U-Net here as φ, with the first half denoted as φ_{down} and the latter half as φ_{up}. Then, we have:

$$\varphi(X_t) = \begin{cases} M'_t, S'_t \text{exists} \& S'_t \geqslant \beta \\ [M_t, M_{t+1}, \cdots, M_{t+L-1}], \text{Otherwise} \end{cases} \tag{2}$$

where M_t represents the predicted human mask at timestamp t, and $M_t\prime$ signifies the human mask of this timestamp previously predicted. The confidence scores of M_t and $M_t\prime$ are denoted as S_t and $S_t\prime$, respectively, where:

$$[S_t, \cdots, S_{t+L-1}] = MLP[P[Conv_1[\varphi_{down}(X_t)]] + P[Conv_2[\varphi(X_t)]]] \tag{3}$$

when the corresponding $S_t\prime$ does not exist. Here, P represents adaptive pooling layers, Conv denotes convolutional layers, and MLP signifies a multi-layer perceptron.

A significant challenge in training the mask prediction network with an end-to-end approach is the absence of a realistic neuromorphic camera dataset containing labeled human body mask sequences for recorded event streams. Nevertheless, as our pro-

posed motion-to-event simulator (outlined in Section 5) can generate paired pixel-level human masks at a high frame rate, they can be utilized to comprehensively train the mask prediction network.

4.3. Human Pose Estimation Network

The human pose estimation network comprises a ResNet-based feature extractor, a Bidirectional Convolutional LSTM (BiConvLSTM) layer, an HPE backbone, and a triangulation module. The initial portion of ResNet34 functions as the feature extractor, succeeded by a BiConvLSTM layer incorporating a skip connection. Given that the human body undergoes no abrupt changes within a relatively brief temporal window, adjacent frames typically exhibit similar ground truth labels. This continuity in human joint movements renders it advantageous to reference neighboring frames when estimating joint positions.

BiConvLSTM represents a bidirectional variant of ConvLSTM [76], in which ConvLSTM constitutes a form of recurrent neural network designed for spatiotemporal prediction, incorporating convolutional structures within both input-to-state and state-to-state transitions. Subsequently, the HPE backbone comprises three hourglass-like CNN blocks, each producing a series of marginal heatmaps to reconstruct the coordinates of human joints in 3D space. All intermediate outputs from these three blocks are utilized to compute the loss with the ground truth heatmaps, while the final two blocks can be regarded as refinement networks. The feature extractor and backbone network architecture have been developed based on a model proposed in [11]. Denoting the feature extractor as γ_1 and the generated feature map as F_t, we have:

$$F_t = \gamma_1(X_t \cdot M_t) \tag{4}$$

where the symbol $\cdot$ denotes pixel-wise multiplication. Consequently, the generated joint heatmaps H_t can be derived as follows:

$$H_t = \gamma_2[Conv[BiConvLSTM(F_t), F_t]] \tag{5}$$

For each joint, YeLan generates three heatmaps depicting the probability of its projected position on the xy, xz, and yz planes (denoted as $H_t^{xy}, H_t^{xz}, H_t^{yz}$). Subsequently, a soft-argmax operator is applied to extract the normalized coordinates of the joint. Ultimately, predictions from the xy plane serve as the final estimations for x and y coordinates, while values for z are computed by averaging the yz and xz predictions. The formula can be expressed as follows:

$$\left[x_{tj}^{xy}, y_{tj}^{xy}\right] = \sigma\left(H_{tj}^{xy}\right) \tag{6}$$

$$\left[x_{tj}^{xz}, z_{tj}^{xz}\right] = \sigma\left(H_{tj}^{xz}\right) \tag{7}$$

$$\left[y_{tj}^{yz}, z_{tj}^{yz}\right] = \sigma\left(H_{tj}^{yz}\right) \tag{8}$$

$$p_{tj}^{xyz} = \left[x_{tj}^{xy}, y_{tj}^{xy}, \frac{z_{tj}^{yz} + z_{tj}^{xz}}{2}\right] \tag{9}$$

where x_t^{xy} denotes the estimated x at time t for a specific joint j's predicted xy-plane heatmap, σ means the soft-argmax operator, and p_{xyz} represents the predicted 3D coordinates for joint j at time t.

The ground truth labels employed during training and testing are normalized prior to being input into the network. For a particular joint, we initially project it onto a plane parallel to the camera's image plane, maintaining the same depth as the depth reference. The head joint's depth value serves as this reference. Subsequently, the 3D space within the DVS camera's view is mapped to a cube with a range of $[-1, 1]$. Finally, as the network does not directly predict the 3D coordinates of a joint but instead forecasts its marginal heatmaps, we extract the joints' projections on three orthogonal faces of the normalized space cube to generate the ground truth for marginal heatmaps. The ultimate marginal heatmaps are computed using a Gaussian filter applied to these projection images [11].

4.4. Loss

For the mask prediction network, the loss function comprises three components. The initial component is the Binary Cross Entropy (BCE) loss, computed between the predicted mask series $\widehat{M}$ and the corresponding ground truth masks M. This loss is implemented to ensure the precision of all generated masks. Subsequently, a Mean Square Error (MSE) loss is calculated over the predicted confidence scores S and their ground truth. The ground truth score represents the Mean Absolute Error (MAE) between a predicted mask and its respective ground truth mask. Finally, although the objective is to predict the mask series for the current frame and subsequent frames, they are not uniformly significant. It is imperative to ensure that masks for proximate frames receive greater weighting, particularly for the current frame. Consequently, an additional BCE loss is computed between the predicted mask for the current frame ($\widehat{M}_0$) and its corresponding ground truth M_0. The cumulative loss is:

$$\begin{aligned} Loss_{mask} = {} & BCE\left(M, \widehat{M}\right) + BCE\left(M_0, \widehat{M}_0\right) \\ & + MSE\left(S, MAE\left(M, \widehat{M}\right)\right) \end{aligned} \tag{10}$$

We employed the loss presented in [11] for the human pose estimation network. As the marginal heatmaps can be interpreted as probability distributions of joint locations, Jensen-Shannon Divergence (JSD) can be applied between the predicted heatmaps by each block ($\widehat{H}^i$, where i denotes the block index) and the ground truth heatmaps H on each projection plane (xy, xz, zy). Additionally, a geometrical loss is computed between the reconstructed 3D joint coordinates $\hat{p}xyz$ and their ground truth $pxyz$. The loss for this stage can be expressed as follows:

$$\begin{aligned} Loss_{HPE} = \sum_i \Big(& ||\hat{p}_{xyz}^i - p_{xyz}||_2 + JSD\left(H_{xy}, \widehat{H}_{xy}\right) + JSD\left(H_{xz}, \widehat{H}_{xz}\right) \\ & + JSD\left(H_{zy}, \widehat{H}_{zy}\right)\Big) \end{aligned} \tag{11}$$

5. Synthetic Data Generation with Motion-to-Event Simulator

First and foremost, while some neuromorphic camera-based HPE datasets exist [10], they primarily focus on fixed everyday movements (e.g., walking, jumping, waving hands), causing models trained on them to falter when faced with intricate movements. Dance performances generally encompass swift and complex gestures, which are scarce in a typical everyday gesture dataset such as [10]. Furthermore, these datasets have been collected under optimal lighting conditions with blank or static backgrounds, failing to represent real-world environments. Additionally, the extant dataset suffers from a lack of diversity in participants, motion dynamics, clothing styles, and types of background activities. Concerning clothing, all participants don identical, form-fitting black attire. To address this critical gap, we propose generating synthetic data with a comprehensive motion-to-event simulator.

Regarding simulators, though several existing neuromorphic camera simulators are available [77,78,74,79], they share a common and pivotal issue. Almost all exclusively convert existing images or videos into event streams, rather than creating highly customized event streams tailored to specific research problem requirements. Furthermore, for existing simulators, if the source video lacks human joint labels, the generated event stream is also devoid of labels. Nonetheless, in this work, the simulator we

develop employs a physics-aware rendering system and renders all relevant parameters fully controlled and customized, including lighting, motion, human gender, body shape, skin color, clothing and accessories, background, and scenes. This feature allows us to implement complex dance movements, and all generated data is paired with accurate labels.

5.1. Advantages of Synthetic Data

Numerous advantages of employing a synthetic data generator have been emphasized in recent work [80], with a car driving-based DVS event simulator also proposed [81]. Acquiring real-world data is both costly and time-consuming, as high costs per participant limit sample size, scene scale, and diversity in physical and environmental attributes. The synthetic dataset generation process affords comprehensive control, enabling precise management of a wide array of parameters. The resulting video file can be rendered at an exceptionally high frame rate (FPS), mitigating blur issues stemming from under or over-exposure in dynamic lighting conditions. Finally, the ground truth of human joints' 3D locations can be extracted within the software for various dances and human models.

5.2. Tools used in the simulator

Synthetic data generation consists of RGB dance video rendering, human joints' position extraction, and events generation. MikuMikuDance (MMD), Blender, and V2E are chosen for each step, respectively.

MMD is a freeware animation program that lets users animate and creates 3D animated movies. This software is simple but powerful, with a long history and a big open-source community behind it. Plenty of human models, scenes, and movement data can be easily accessed for free. In addition, it can automatically handle clothing physics and interaction with the body in a sophisticated manner with minimal manual adjustment.

Software *Blender* is adopted to generate human joints' ground truth labels and camera matrix. *Blender* is a free and open-source 3D computer graphics software for creating animated films, motion graphics, etc. It is highly customizable, and all essential information can be accessed during rendering, including the 6 degrees of freedom coordinates of human joints and the camera center. The 13 key points' ground truth coordinates are extracted at 300 frames per second (FPS).

Lastly, *Video to Event (V2E)* is used for event generation. V2E is a toolset released in 2021 by Delbruck et al. [74]. It can synthesize realistic event data from any conventional frame-based video using an accurate pixel model that mimics the neuromorphic camera's nonidealities. According to its author, V2E supports an extensive range of customizable parameters and is currently the only tool to model neuromorphic cameras realistically under low illumination conditions.

5.3. Comprehensive Motion-to-Event Simulator

As Fig. 3 shows, our proposed motion-to-event generator takes 3D character models, motion files, camera views, and lighting conditions as inputs. MMD renders RGB dance videos and their paired human mask videos given these inputs, while Blender generates corresponding ground truths. For each camera view, a camera intrinsic and extrinsic matrix is calculated by Blender as well.

Then our simulator combines these rendered dance videos with collected background videos by referring to the paired mask videos. According to [74], if a video's temporal resolution is low,

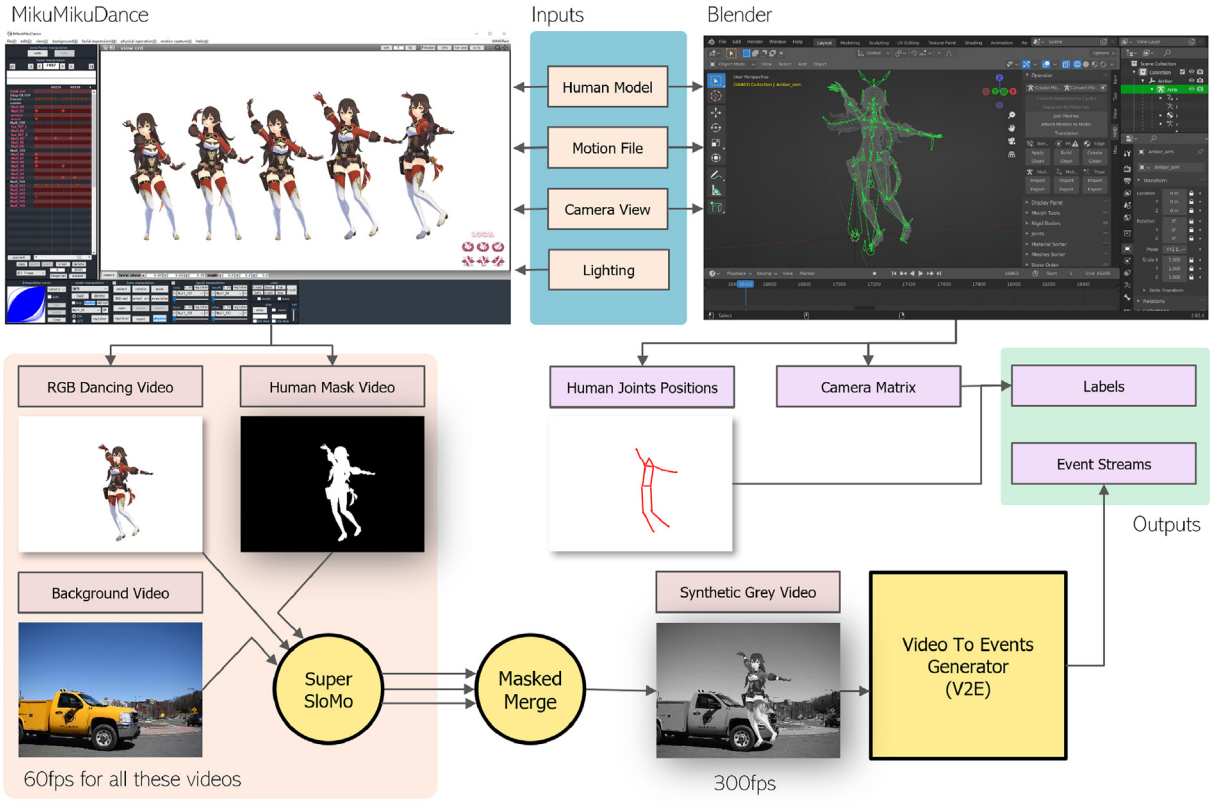

Fig. 3. The pipeline for synthetic data generation utilizing the comprehensive motion-to-event simulator comprises motion files, human models, camera views, lighting, and other settings as input. These inputs are rendered into RGB and human mask videos in MMD, subsequently merged with background videos. The merged videos undergo processing as event streams using V2E, while the corresponding labels are computed in Blender.

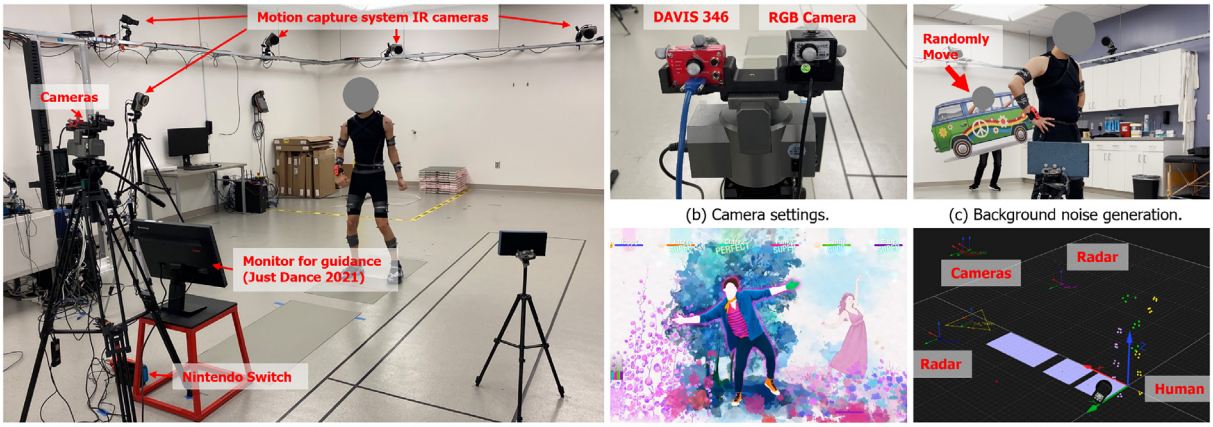

Fig. 4. Settings of human subject data collection in an indoor laboratory setting.

generated event stream will be less realistic. However, due to the software limitation and background video quality, the synthesized videos are at 60 FPS. This gap in FPS is compensated using SuperSlowMo[82], which can interpolate videos to a high FPS with convincing results. To reduce the time and computation cost, we only apply SuperSlowMo to dance and background videos before the merging. This way, we do not have to interpolate synthesized videos for each human and background combination.

After the RGB video rendering, merged videos are sent to the V2E events stream generator. Many parameters such as event trigger thresholds, noise level, and slow-motion interpolation scale, can be modified in detail. These features enable us to generate many highly customizable DVS event streams at a meager cost in a short time. To simulate situations in the real world, we increase the noise as the brightness decreases.

For human joints' ground truth, as mentioned above, we write our customized scripts and inject them into Blender to collect the exact position of all joints while rendering scenes at 300 FPS. Also, scripts help extract the camera's intrinsic and extrinsic matrix used in the label pre-processing. Besides the advantages mentioned above, 3D human models have even more initial flexibility. Skin color, height, body style, clothing, hair color, style, and accessories are all easily modifiable - which is very difficult to do in real-world data collection.

5.4. Synthetic Dataset Description

Utilizing the comprehensive motion-to-event simulator, we have generated a vast synthetic dataset *Yelan-Syn-Dataset*, comprising approximately 4 million data samples (specifically, 3,958,169 frames). Examples of synthetically generated RGB frames are shown in Fig. 5. The total dataset size is around 2.6 terabytes. This data was synthesized from 1320 combinations of various variables, including 10 human models, 8 one-minute dance motions, 11 background videos, 4 camera views (i.e., front, back, left, right), and 3 different lighting conditions (i.e., high, medium, low). This dataset contains processed TORE volume, paired labels and masks, constant-time frames with identical time steps

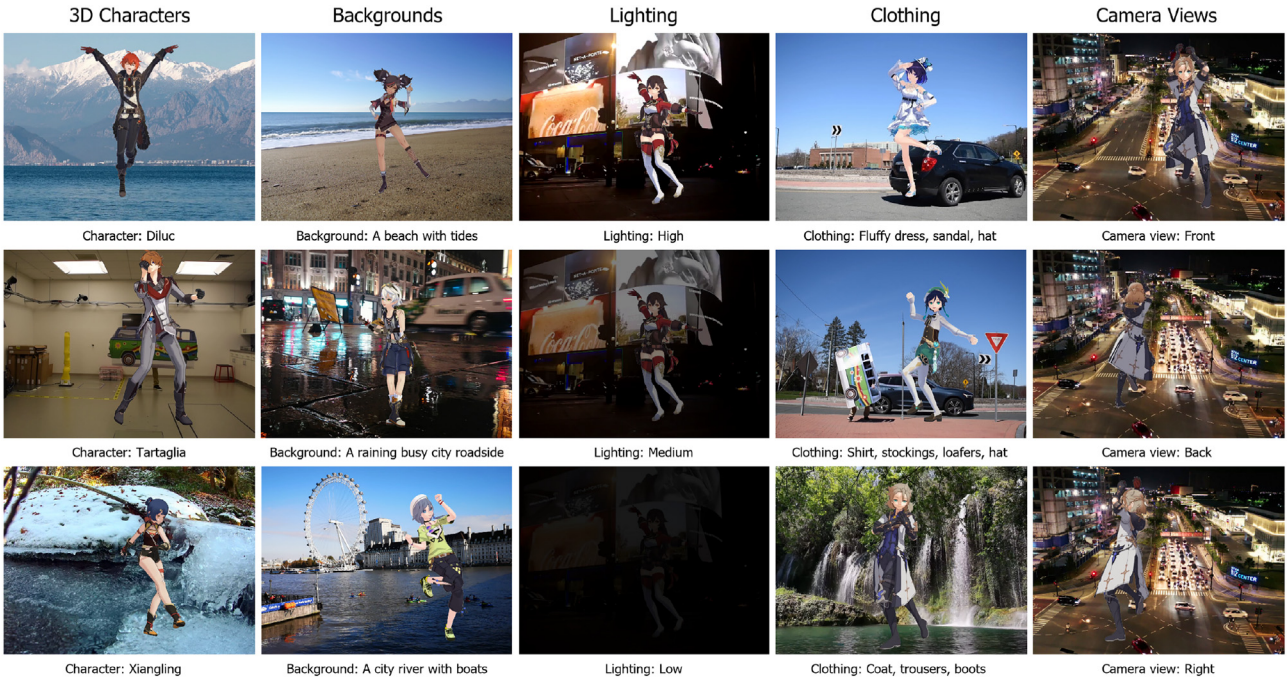

Fig. 5. RGB frame samples from the synthetic data generated by the comprehensive motion-to-event simulator.

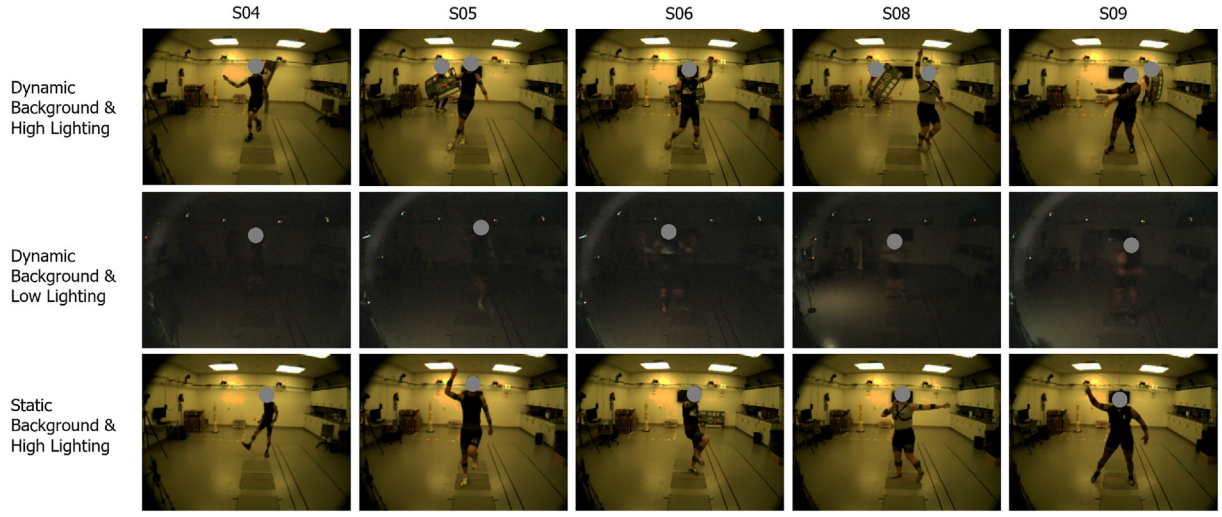

Fig. 6. RGB frame samples from the *Yelan-Real-Dataset* collected in motion capture laboratory.

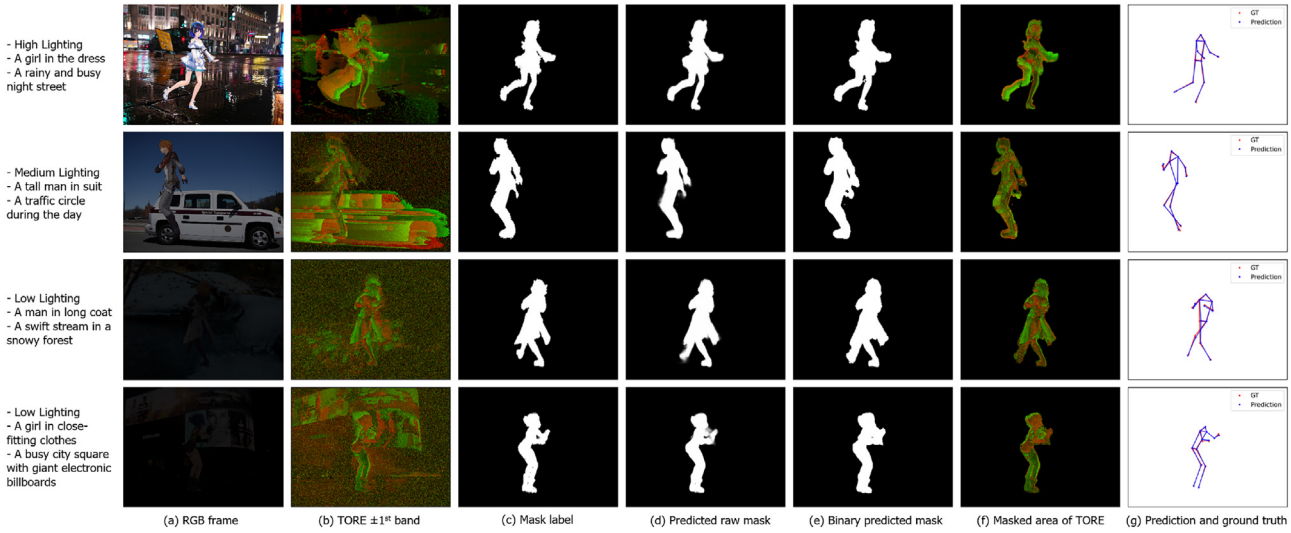

Fig. 7. Sample results from *Yelan-Syn-Dataset* in different lighting conditions with dynamic backgrounds. The RGB frames, corresponding event representation TORE, generated masks, ground truth, and predicted 3D human pose are shown.

(20 ms), and raw event stream files for generating any other customized representations. Due to computational resource and training time constraints, only 30% of data instances from 330 selected setting combinations are employed for training and validation. They are shuffled and partitioned with an 8:2 ratio. Subsequently, we randomly select 82 setting combinations from the remaining 990 unused combinations as the test set. (See Fig. 6).

6. Human Subject Study with Dynamic Lighting and Background

6.1. Real-World Dataset in Motion Capture Facility

In addition to the synthetic dataset *Yelan-Syn-Dataset*, we have conducted an Institutional Review Board (IRB)-approved human subject study to collect a real-world human pose dataset, Yelan-Real-Dataset, from nine participants. Seven of the participants are male and two are female. The participants, aged 20–31, were recruited from a university campus using a snowball sampling technique. During this study, our participants were asked to play the Nintendo dance game "Just Dance 2021". Each participant selected five songs to dance to during the study after a brief training period with the tutorial dances in the Nintendo dance game. A monitor provided further guidance/cues for participants to follow along. (See Fig. 7).

A 9-camera-based 3D motion capture system (Qualisys AB, Göteborg, Sweden) was used to obtain ground truth 3D kinematics data of the human body, providing the 3D absolute coordinates for all selected human joints at a frame rate of 200 frames per second (FPS). A neuromorphic camera (DAVIS 346 [83]) and an RGB camera were simultaneously operated to record the participants' movements. Other essential equipment included a flicker-free LED light, IR filter, cardboard background, monitor, and Nintendo Switch. During the dataset collection, the lighting condition was strictly controlled to induce low-lighting and high-lighting conditions. All the lights except a dim flicker-free lamp were turned off during the low-lighting conditions session. The IR filter was attached in front of the neuromorphic camera's lens to filter out the events caused by IR light emitted by the motion capture

system. Fig. 4 illustrates the data collection settings, hardware, sensor arrangements, and mechanisms to generate background noise.

In the dynamic background condition, a person behind the target participant randomly moved with a large cardboard vehicle. On the other hand, for static background cases, no other movable contents appear in the background during recording. On average, each participant danced for about 20 min, and the time was distributed equally to the four conditions mentioned above. Static background cases are only included for result comparison, as mask-based background filtering is not helpful in these cases. We divided *Yelan-Real-Dataset* simply by selecting all the events generated by participants seven (male) and eight (female) as test sets. All other data is shuffled and divided into training and validation sets with the same 8:2 ratio. We prepared a short video to further illustrate the rich synthetic and real-world data, which can be found at bit.ly/yelan-research.

7. Results

7.1. Training Details

In this study, all models are trained using eight 1080ti or 2080ti graphics cards. The batch size remains consistent throughout the training process: stage one employs a batch size of 128, while stage two utilizes a batch size of 16. Training parameters for all stages include a learning rate of 0.001, an Adam optimizer with a weight decay of $1e-5$, and early stopping with a patience of 10 epochs. A learning rate scheduler is implemented to reduce the learning rate by half every N epochs, with N set to 5 in stage one and 10 in stage two to account for differences in training patterns and convergence speeds. This project is executed using Pytorch-Lightning [84]. *YeLan*'s stages one and two are trained separately. Owing to ground truth limitations, the mask prediction network is trained exclusively on synthetic data and applied directly in both *Yelan-Syn-Dataset* and *Yelan-Real-Dataset*. Stage two is initially pretrained on *Yelan-Syn-Dataset* and subsequently fine-tuned on *Yelan-Real-Dataset*.

7.2. Evaluation Metrics

In accordance with convention, the primary evaluation metric employed in this study is the Mean Per-Joint Position Error (MPJPE), a widely used metric in HPE. Additionally, two other popular metrics, PCK and AUC, are considered in the main comparison. The MPJPE is calculated by averaging the Euclidean distance between each predicted joint and its corresponding ground-truth joint, typically measured in millimeters. PCK represents the Percentage of Correct Keypoints according to a commonly used threshold value of 150 mm. A predicted joint is considered correct and given a value of 1 if it is within a 150 mm cube centered on the ground truth joint; otherwise, it is deemed incorrect and assigned a value of 0. AUC refers to the Area Under the Curve for the PCK metric at varying thresholds. Standard threshold sets, consisting of 30 evenly spaced numbers from 0 to 500 mm, are used in this study. The target AUC is determined by first calculating the PCK at all threshold values and then computing their mean value.

$$MPJPE = \frac{1}{J}\sum_{i}^{J}||p_{xyz} - \hat{p}_{xyz}|| \tag{12}$$

$$PCK_{\alpha} = \frac{1}{J}\sum_{i}^{J} sign(\alpha - ||p_{xyz} - \hat{p}_{xyz}||) \tag{13}$$

$$AUC = \frac{1}{N}\sum_{\alpha=0}^{A} PCK_{\alpha}(p_{xyz}, \hat{p}_{xyz}) \tag{14}$$

joint positions, respectively, while J represents the number of skeleton joints. For PCK and AUC calculations, α corresponds to the threshold and A refers to the maximum threshold. N in AUC indicates the number of different thresholds utilized.

In the present study, we selected baseline methods introduced by Scarpellini et al. [11] and Baldwin et al. [73], both initially published in 2021. Scarpellini et al. employed two event-count-based representations–constant event count frames and voxel grid–that exhibit a variable time step, causing a loss of synchronization with our labels and TORE volumes. As a result, we adapted the original work to accept constant time representations as input, setting the time step for this representation at 20 ms to ensure accurate label matching.

Baldwin et al. presented the TORE volume and evaluated it using the model and dataset proposed by Calabrese et al. [10], employing an alternative representation to demonstrate the superior performance of their novel representation in the multi-view human pose estimation task. With the goal of developing a monocular human pose estimation framework in this paper, we utilized the same monocular HPE model from Scarpellini et al. [11], substituting the representation with TORE to showcase the efficacy of Baldwin et al.'s [73] method in this new task.

Alongside the hourglass-like HPE backbones selected for this study, we also explored other commonly used structures, such as residual block-based and bottleneck block-based backbones [85]. Throughout these experiments, we maintained a consistent number of blocks for all designs and kept all other settings and conditions constant.

7.3. Evaluation

Table 2 illustrates the performance comparison between the proposed *YeLan* and baseline models on both *Yelan-Syn-Dataset* and *Yelan-Real-Dataset*. The table shows that the baseline models consistently underperform the proposed YeLan model in both conditions. In *Yelan-Syn-Dataset*, the model proposed by the Scarpellini et al. [11] achieves an MPJPE and PCK of respectively 91.88 and 83.92% while the model proposed by the Baldwin et al. [73] with TORE representations have a slightly improved performance of respectively 59.34 MPJPE and 93.17% PCK. The proposed *YeLan* system outperforms the baseline models by a significant margin and achieves the lowest MPJPE of 46.57 and the highest PCK of 96.34% on *Yelan-Syn-Dataset*. (See Table 3).

Compared to the synthetic data, the performance of all the models on the real-world data is worse. The real-world data comes with its own challenges and unique characteristics, like more noise and limited transmission bandwidth. Compared to the perfect ground truth labels generated in the synthetic data, some non-negligible spatial and temporal mismatch happens in the real-world data, especially considering the fact that the background activities often block some IR cameras used for motion capture. There is also substantial heterogeneity in the skeletal formations and impedance-matching states in the real-world data. Lastly, in the real-world data, another researcher generating background activities are also in the camera's view, who is sometimes recognized and masked as an additional human and confuse the models as a result. Consequently, the performance of the real-world data will be expected to be lower than that of the synthetic data. However, the proposed *YeLan* still achieves the lowest metrics compared to the baseline models. The baseline models are designed to work in ideal environments which do not suffer from dynamic lighting conditions and moving background content problems, which contributes to their poor performance in more realistic environments.

Table 2
Test set results comparison on three metrics for *Yelan-Syn-Dataset* and *Yelan-Real-Dataset*. †: P means YeLan with pre-training on Yelan-Syn-Dataset and tuned on Yelan-Real-Dataset. HG: Hourglass-like blocks backbone. RES: Residual backbone. BTN: Bottleneck backbone.

		Yelan-Syn-Dataset			Yelan-Real-Dataset		
Methods	Representation	MPJPE↓	PCK(%)↑	AUC(%)↑	MPJPE	PCK	AUC
Scarpellini et al., 2021	Constant Time	91.88	83.92	80.95	111.57	78.12	77.03
Baldwin et al., 2021	TORE	59.34	93.17	86.97	101.97	82.22	78.90
YeLan - BTN	TORE	126.50	84.70	81.49	137.21	69.74	72.47
YeLan - RES	TORE	58.95	93.51	86.97	112.49	77.76	76.88
YeLan - HG	TORE	**46.57**	**96.34**	**89.37**	96.61	81.88	79.78
YeLan - HG - P†	TORE	-	-	-	**90.94**	**85.14**	**80.91**

Table 3
The comparison between RGB, RGBD, and a neuromorphic camera, DAVIS346. For FOV, H, V, and D stand for horizontal, vertical, and diagonal FOV.

Camera Type	RGB (In DAVIS 346)	RGBD (Intel RealSense 435i)	DVS (DAVIS346)
Update Rate	Up to 40 Frames/ sec	30 Frames/ sec	Up to 12 MEvents/ sec
FOV	Any large, Dependent on lens. Lens are replaceable.	IR Projector: Depth: H: 90 ± 3 H: 87 V: 63 ± 3 V: 58 D: 99 ± 3 D: 95 FOVs are fixed.	Any large, Dependent on lens. Lens are replaceable.
Dynamic Range	56.7 dB	N/A	120 dB
Power Consumption	140mW	Maximum Power: 2850mW (Measured on Windows 10)	10-30mW (activity dependent)
Working Distance	N/A	0.2∼3 m, varies with lighting conditions.	N/A

7.4. Result Statistics and Analysis

As Fig. 8 shows, *YeLan* has strong stability across the changes in lighting conditions and camera views, while more complicated background contents cast an impact on the results. Different human models also show an impact on performance. From the performance statistics, we observe that the system performs less well for human models "Albedo", "Seele", and "Xiangling". By looking at the original 3D model and test data composition, the reason becomes obvious: Albedo wears a long coat reaching his knee, while Seele wears a fluffy dress with a complex structure. These factors make their joint position harder to estimate. As for Xiangling, though she does not have clothing-related problems, we find some body deformation that happened during the simulation. Also, the randomly selected test cases for Xiangling contain more difficult backgrounds and lighting combinations, such as two busy city street environments in low-lighting conditions, according to Fig. 8 (a).

Moreover, Fig. 9 shows that our model behaves better consistently in all lighting conditions compared to baseline models. This feature is significant in real-world applications as the pipeline could generate convincing predictions regardless of the lighting conditions.

The mask prediction network in *YeLan* enables it to reach a similar level of PCK and AUC whether the background is dynamic or static. We compared the real-world test set with dynamic and static background cases using the model pretrained on the *Yelan-Syn-Dataset* and fine-tuned it on *Yelan-Real-Dataset*. Fig. 10 depicts that *YeLan* get a comparable result in both scenarios.

7.5. Impact of Pretraining on Synthetic Data

Synthetic data from the physics-based simulator is a core contribution of this work which allows us to generate event streams from diverse virtual 3D characters in different simulated settings. They include different clothing, lighting conditions, camera viewing angles, dynamic backgrounds, and movement sequences. The generated synthetic data is faithful to physics and can provide essential data for pretraining *YeLan* since running real-world experiments with a large number of participants is expensive and time-consuming. To this end, we hypothesize that the pretraining on synthetic data will allow *YeLan* to learn better feature representation of physical conditions in diverse simulated settings and achieve superior performance on the *Yelan-Real-Dataset*.

Table 2 demonstrates that our proposed model achieves superior performance (comparing MPJPE/PCK/AUC metrics) on the

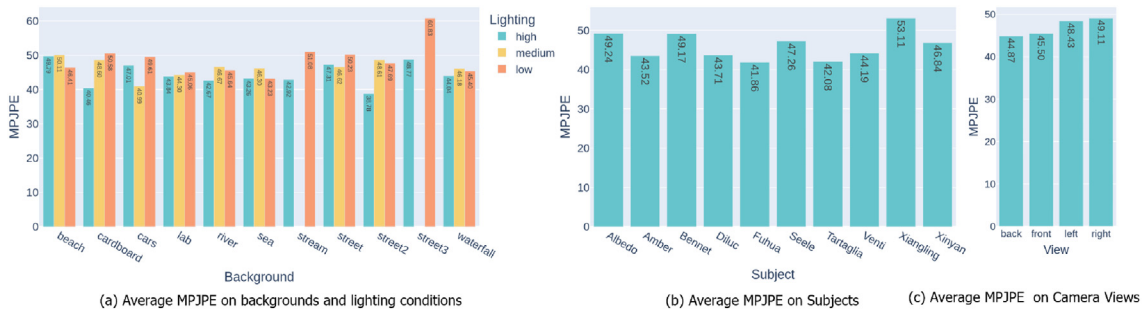

Fig. 8. The performance across lighting conditions, dynamic backgrounds, subjects, and camera views in our synthetic data.

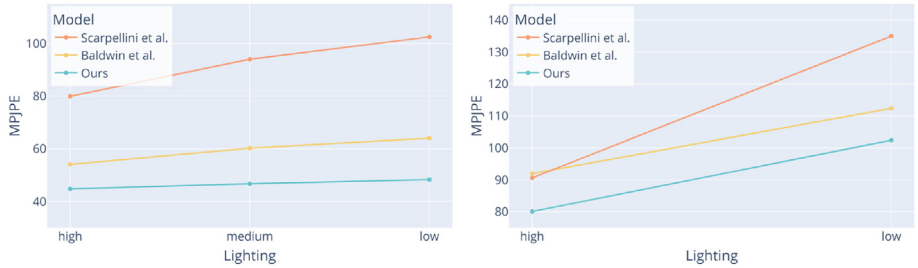

Fig. 9. Impact of lighting condition on the performance (MPJPE) on our synthetic (left) and real-world (right) dataset.

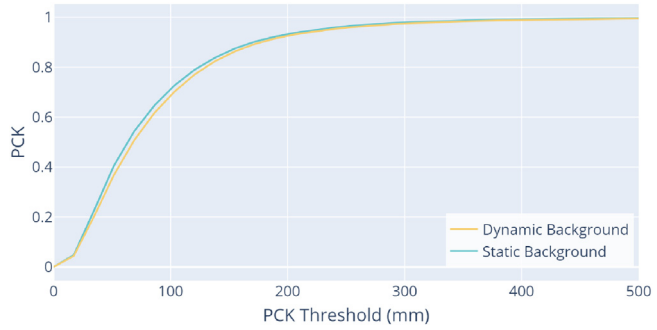

Fig. 10. AUC curve with dynamic and static background on *Yelan-Real-Dataset*.

Yelan-Real-Dataset if the model is pretrained on *Yelan-Syn-Dataset* from the simulator. The model trained only on the *Yelan-Real-Dataset* achieves 96.61 MPJPE and 81.88% PCK. On the other hand, if the model is pretrained on *Yelan-Syn-Dataset* and then fine-tuned on the *Yelan-Real-Dataset*, it achieves 90.94 MPJPE and 85.14% PCK, which clearly shows the benefit of pretraining on synthetic data from a simulator.

7.6. Ablation Study

As introduced in Section 4, there are two most essential modules in *YeLan*: a mask prediction network and the BiConvLSTM. We do an ablation study by removing one module each time and comparing their performance on *Yelan-Syn-Dataset*. If the mask prediction network is removed, the resulting MPJPE is 63.99; if the BiConvLSTM is removed, the corresponding MPJPE is 49.08. In contrast, the complete YeLan pipeline gets an MPJPE of 46.57, which proves the effectiveness of these modules.

7.7. Occlusion

In the real world, occlusion is also an inevitable problem. In *Yelan-Syn-Dataset*, different types of clothes and accessories like hats, long hair, fluffy skirts, and long coats already occluded the human body to a relatively large extent. Moreover, the camera's side views also introduced many self-occlusion. The excellent performance over all these scenarios shows a solid ability to survive occlusion.

To further prove the occlusion resiliency of the system, we augmented the dataset with more block occlusions. In the eye of the neuromorphic camera, if a static object is placed in front of a human and shadows him, the corresponding area shoots no event due to the object's immobility. Regarding the input TORE volume, the occluded area becomes pure black, as no event activity is recorded. To simulate the occlusion like this, we trained a model with the value of random areas set to zero. These rectangular occlusion areas have random sizes and locations, and the occlusion is also applied randomly with a probability of 80% during the training. When testing on the test set with random occlusion enabled, the MPJPE is 96.794. During the test, the occlusion probability is set to 100%. There is an accuracy drop, but considering that the maximum random occlusion area is 80×80 (where the frame size is 260×346), it makes sense as sometimes the majority of the human body could be occluded. Fig. 11 shows samples from the test set.

7.8. Systems Benchmarking

As is introduced in 4.2, the first stage of *YeLan* is an early-exit-style human mask prediction network, where a threshold β is used to decide when to start a new inference. The selection of β is a trade-off. If the threshold is set too high, the mask prediction process will be executed too many times, resulting in a longer inference time. On the contrary, many defective masks will be used if the threshold is too low, which harms the overall accuracy. In order to select the best threshold, we ran an experiment to do human pose estimation on a continuous one-minute event stream with a series of different thresholds. The accuracy and time consumption are recorded and made into Fig. 12. From this figure, we can observe that as we increase the threshold, accuracy goes up and time consumption goes down, while we can achieve a balanced accuracy and time consumption at near 0.99.

Furthermore, to comprehend the error distribution among various joints, we computed the mean MPJPE for all 13 joints. As illustrated in Fig. 13, the head joint exhibits the lowest MPJPE, while the shoulder joints possess the second-lowest error. Conversely, the left and right hands exhibit the highest MPJPE, with foot joints following closely. Our observations indicate that this occurs because the head and shoulders maintain the most stable contours across distinct characters and exhibit minimal overlap with other body parts. Nevertheless, as hands and feet are situated farthest from the human center and exhibit the most movement during dancing, they prove more challenging to estimate due to an increased range of potential movements and patterns. Greater pattern variability contributes to increased uncertainty, which further undermines accuracy when occlusion occurs.

The total model size of *YeLan* amounts to 234 MB, with the total parameter count reaching 61.2 M, of which the stage one model comprises 40.8 M and stage two encompasses 20.4 M. The total FLOP stands at 1482G. In terms of time consumption, we conducted a comprehensive test on the MiDHP test set using a batch size of one, resulting in an average time expenditure of approximately 29.1 ms per frame when executed on a 2080ti graphics card.

8. Comparison between Different Modalities

8.1. RGB Camera

In order to show the superiority of DVS in low-lighting conditions, we also compare our result with RGB frame-based human

Fig. 11. Samples of occluded inputs, predicted mask, TORE volume representation, and inferred human poses.

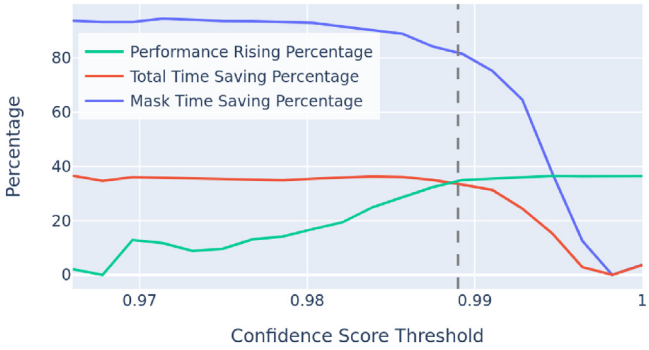

Fig. 12. Time-saving and performance rising percentage with regard to confidence score threshold.

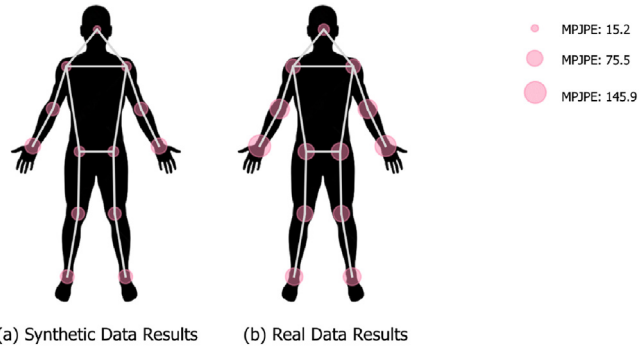

Fig. 13. Joint-wise average MPJPE analysis on *Yelan-Syn-Dataset*.

pose estimation algorithm. We use the DVS camera DAVIS346, where both events and RGB frames are synchronously recorded. As the same device records both modalities via the same lens, there is no difference in the quantity of light captured by RGB and DVS. OpenPose [86] proposed and implemented by Cao et al. is adopted as our RGB baseline. OpenPose is an accurate, fast, and robust 2D human pose estimation algorithm. Although there are differences in joint number and dimension, we calculated the 2D projection of 3D joints generated by *YeLan*. We picked the closest 13 joints from all 25 OpenPose output joints to compare. The comparison is conducted on paired DVS and RGB recordings from the dataset collected in the motion capture lab. Both high and low-lighting

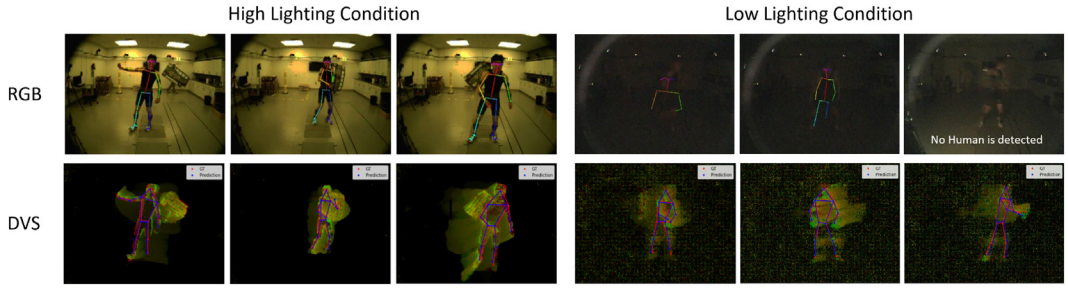

Fig. 14. HPE comparison between RGB and DVS in both high and low-lighting conditions. RGB and DVS are synchronized.

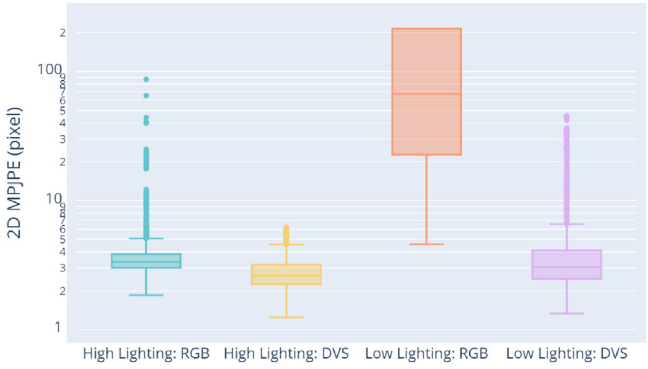

Fig. 15. 2D MPJPE Comparison between DVS and RGB.

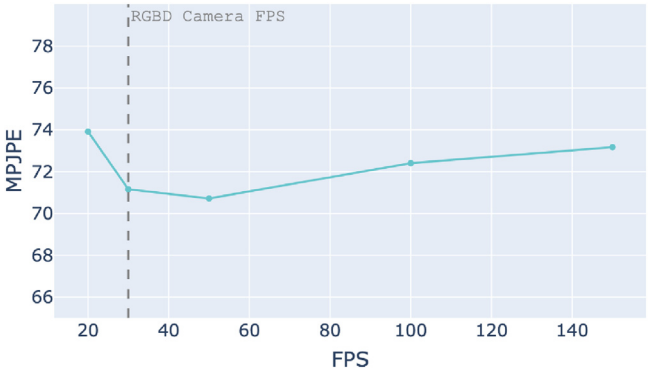

Fig. 16. Stable human pose estimation provided by YeLan on different FPS settings. The model used is only trained on the 50FPS setting. On the contrary, the depth camera has a fixed 30FPS frame rate.

condition cases are included, and qualitative comparison is shown in Fig. 14. Furthermore, we made a box plot showing the 2D MPJPE of DVS and RGB in two lighting conditions in 15. From this figure, it is evident that *YeLan* generates better and more stable predictions in all cases. Although the RGB-based method achieves good performance (marked by the low 2D MPJPE) in high-lighting conditions, the performance deteriorates significantly in the low lighting conditions. OpenPose fails to detect any human from frames due to motion blur and low SNR sensor data when the illumination is lower than a certain threshold. On the other hand, *YeLan* shows strong robustness against lighting conditions changes and constantly generates high-quality predictions (marked by the low MPJPE in both cases).

8.2. RGB-Depth Camera

Besides the RGB camera, the depth camera is also widely used in digital dancing games. [87,88] These cameras are usually paired with RGB cameras, which grant them information from both domains. Compared to RGB cameras, these cameras have a better understanding of the 3D space, which results in a more accurate human segmentation and joint location estimation. RGBD camera-based human pose estimation is a well-established problem with many commercial products and pipelines [89–91], like the Kinect from Microsoft and the RealSense from Intel.

However, the depth camera also has several disadvantages. Depth cameras actively emit and recapture the reflected infrared to build the 3D point cloud. This procedure is significantly more power-hungry (about 200 times higher power consumption than the neuromorphic camera) and suffers from many limitations. Firstly, due to the manufacturing and power consumption consideration, depth cameras' field of view (FOV) is usually fixed and small.

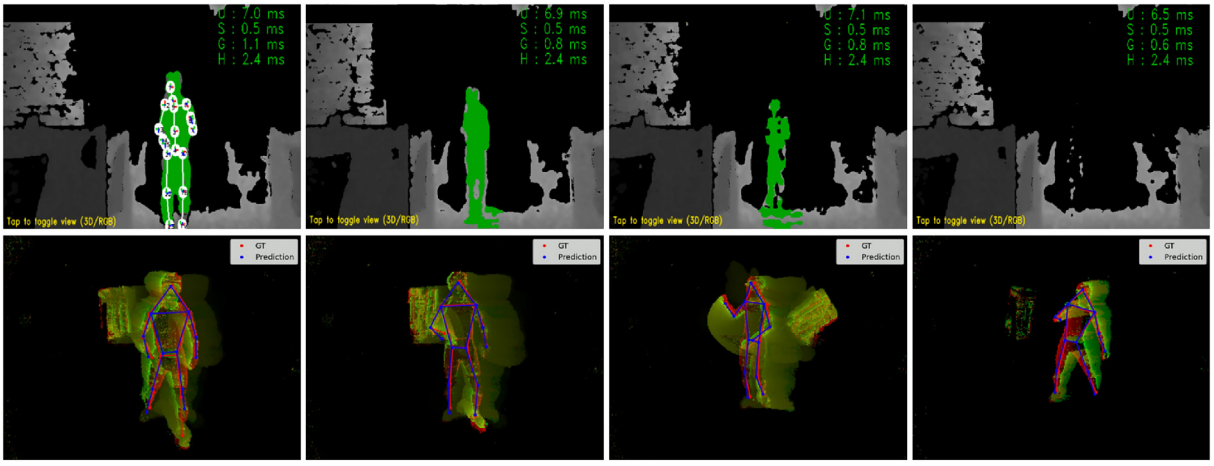

Fig. 17. Qualitative comparison between the RGB-Depth camera solution and YeLan in terms of the distance between the camera and the human subject. The first row displays results from the depth camera, while the second row presents outcomes from YeLan. The distance increases from left to right.

Small FOV limits the sensing space coverage and restricts the dancer's movement in a smaller space. Secondly, the distance between the target and the depth camera strongly impacts the detection accuracy. If the sensor is too close to the target (smaller than 0.5 m), the two IR receivers have overlapped IR patterns which saturate the IR camera and result in an estimation failure [92]. Moreover, when the distance is far, the detection rate drops drastically after a certain point, as the IR receiver cannot receive enough reflected IR light. The first two points together limit the functional dancing area and cause a substantial restriction on relative distance.

Thirdly, since the camera emits light itself, it could also be interfered with by other light sources, which is especially true for strong light sources like stage lights and solar lights. This characteristic restricts the application of depth cameras in outdoor and indoor spaces near the light source.

Fourthly, as we have previously mentioned, our system is able to work at any high FPS, and the only procedure needed to take is to modify the time step size when generating the TORE volumes for an event stream. Compared with the neuromorphic camera, the maximum RGBD camera FPS is usually very low, making it hard to be applied to track fast-paced dances or generate high-fidelity 3D digital dances. The two most commonly used RGBD cameras, i.e., Microsoft Azure Kinect and Intel RealSense, support at most 30 FPS when capturing full RGBD streams. On the contrary, neuromorphic cameras can easily receive ten-millions level events per second, which gives them overwhelmingly huge advantages over RGBD cameras. We tested our model by generating representations from 20FPS to 150FPS (due to the restriction of ground truth label rate) on all the event streams from test subject eight. The result shows that the estimation accuracy is pretty stable on various frame rate settings (Fig. 16).

Although it would be ideal to quantitatively compare RGBD and neuromorphic cameras for human pose estimation, these two types of cameras can not work simultaneously. When the depth camera is turned on, its built-in IR projector emits a grid of IR rays 30 times per second. In the eye of the neuromorphic camera, the IR projection of the depth camera is visible and everything in the surrounding is constantly flickering at a high frequency with meshed dots, which negatively impacts the performance. Therefore, we made a qualitative comparison by separately recording two sessions of human dance by Intel RealSense 435i and DAVIS 346. During these two sessions, the participant moves from 2.5 m to 3.5 m with some dynamic dance patterns, and the results are shown in Fig. 17. As the figure shows, as the distance between the camera and the human increases, the RGB-Depth camera gradually fails to capture valid human shapes. The human pose estimation program also stops generating valid predictions after a certain point. For the RGB-Depth camera-based 3D human pose estimation, we chose the Nuitrack [93], a commercial product designed specifically for RGBD camera-based HPE problems, for comparison. Although RGB and depth channels are all used, here, for visualization convenience, only the depth channel is shown, and the estimated human masks are shown in green.

9. Limitation and Future Works

Overall, YeLan clearly highlights the promise of the 3D human pose estimation in the context of dancing in the presence of different confounding factors. However, the present study exhibits several limitations that we intend to systematically address in future work. In the current work, one assumption was that the neuromorphic camera is stationary and attached to a fixed tripod. Consequently, YeLan is not readily transferable to scenarios where the neuromorphic camera can move and is mounted on a car or a drone. The current version of YeLan has been primarily tested in a single-person scenario and does not support a multiplayer dancing game. By incorporating multi-person segmentation and masking, we plan to extend YeLan for multi-person situations. Additionally, we aim to investigate energy-efficient implementations on neuromorphic computing platforms such as Intel Loihi 2. [94].

10. Conclusion

This work discussed existing 3D HPE techniques employed in dance games and evaluated their strengths and limitations. We proposed an innovative neuromorphic camera-based approach to address these shortcomings. We amassed a real-world dance dataset through human subject studies and constructed an extensive motion-to-event simulator to generate a vast amount of fully controllable, customizable, and labeled synthetic dance data to aid in pre-training the model. YeLan surpasses all baseline models in various challenging scenarios on both datasets. Furthermore, we conducted an in-depth analysis and comparison between different modalities, which unequivocally demonstrates YeLan's superiority in many aspects.

CRediT authorship contribution statement

Zhongyang Zhang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Kaidong Chai:** Software, Data curation, Investigation. **Haowen Yu:** Software. **Ramzi Majaj:** Resources, Investigation. **Francesca Walsh:** Methodology, Writing - review & editing. **Edward Wang:** Supervision, Writing - review & editing. **Upal Mahbub:** Supervision. **Hava Siegelmann:** Supervision, Funding acquisition. **Donghyun Kim:** Supervision, Resources. **Tauhidur Rahman:** Supervision, Project administration, Funding acquisition, Resources, Writing - original draft.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hava Siegelmann reports financial support was provided by Defense Advanced Research Projects Agency. Tauhidur Rahman reports financial support was provided by Defense Advanced Research Projects Agency. Tauhidur Rahman reports financial support was provided by National Science Foundation. Coauthor employed by Qualcomm Technologies, Inc. - Upal Mahbub..

Acknowledgments

We express our gratitude to the Institute of Applied Life Sciences and the College of Information and Computer Sciences at UMass Amherst for supplying start-up funds and laboratory support. Additionally, we extend our appreciation to the HDSI department at UCSD for their startup funding contributions. This work was also partially financed by the DARPA TAMI grant (Project ID HR00112190041), a grant from Qualcomm Technologies, Inc., and the directorate for computer science and engineering of the NSF (Award Number 2124282). Finally, we offer special thanks to miHoYo Co., Ltd. for providing high-quality character models utilized in synthetic data generation.

References

[1] M.M. López-Rodríguez, M. Fernández-Martínez, G.A. Matarán-Peñarrocha, M.E. Rodríguez-Ferrer, G.G. Gámez, E.A. Ferrándiz, Efectividad de la biodanza

acuática sobre la calidad del sueño, la ansiedad y otros síntomas en pacientes con fibromialgia, Medicina Clínica 141 (11) (2013) 471–478.
[2] S.-L. Cheng, H.-F. Sun, M.-L. Yeh, Effects of an 8-week aerobic dance program on health-related fitness in patients with schizophrenia, Journal of Nursing Research 25 (6) (2017) 429–435.
[3] D.X. Marquez, R. Wilson, S. Aguiñaga, P. Vásquez, L. Fogg, Z. Yang, J. Wilbur, S. Hughes, C. Spanbauer, Regular latin dancing and health education may improve cognition of late middle-aged and older latinos, Journal of aging and physical activity 25 (3) (2017) 482–489.
[4] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, J. Luo, Anatomy-aware 3d human pose estimation with bone-based pose decomposition, IEEE Transactions on Circuits and Systems for Video Technology 32 (2022) 198–209.
[5] Y. Chen, Y. Tian, M. He, Monocular human pose estimation: A survey of deep learning-based methods, Comput. Vis. Image Underst. 192 (2020).
[6] M. Hassan, V. Choutas, D. Tzionas, M.J. Black, Resolving 3d human pose ambiguities with 3d scene constraints, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 2282–2292.
[7] P. Lichtsteiner, C. Posch, T. Delbruck, A 128 ×128 120 db 15 mus latency asynchronous temporal contrast vision sensor, IEEE journal of solid-state circuits 43 (2) (2008) 566–576.
[8] C. Posch, D. Matolin, R. Wohlgenannt, An asynchronous time-based image sensor (2008) 2130–2133 doi:10.1109/ISCAS.2008.4541871.
[9] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A.J. Davison, J. Conradt, K. Daniilidis, et al., Event-based vision: A survey, IEEE transactions on pattern analysis and machine intelligence 44 (1) (2020) 154–180.
[10] E. Calabrese, G. Taverni, C.A. Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, T. Delbrück, Dhp19: Dynamic vision sensor 3d human pose dataset, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 1695–1704.
[11] G. Scarpellini, P. Morerio, A. Del Bue, Lifting monocular events to 3d human poses, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1358–1368.
[12] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE transactions on pattern analysis and machine intelligence 36 (7) (2013) 1325–1339.
[13] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, T. Brox, 3d human pose estimation in rgbd images for robotic task learning, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 1986–1992.
[14] S. Zou, C. Guo, X. Zuo, S. Wang, P. Wang, X. Hu, S. Chen, M. Gong, L. Cheng, Eventhpe: Event-based 3d human pose and shape estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10996–11005.
[15] M.F.H. Sánchez, E.D.B. Marín, L.T.O. Mora, Characterization of dance-based protocols used in rehabilitation-a systematic review, Heliyon (2021).
[16] L. Teixeira-Machado, R.M. Arida, J. de Jesus Mari, Dance for neuroplasticity: A descriptive systematic review, Neuroscience & Biobehavioral Reviews 96 (2019) 232–240.
[17] M. Akandere, B. Demir, The effect of dance over depression, Collegium antropologicum 35 (3) (2011) 651–656.
[18] H. Hashimoto, S. Takabatake, H. Miyaguchi, H. Nakanishi, Y. Naitou, Effects of dance on motor functions, cognitive functions, and mental symptoms of parkinson's disease: a quasi-randomized pilot trial, Complementary therapies in medicine 23 (2) (2015) 210–219.
[19] M. del Mar López-Rodríguez, A.M. Castro-Sánchez, M. Fernández-Martínez, G. A. Matarán-Penarrocha, M.E. Rodríguez-Ferrer, Comparación entre biodanza en medio acuático y stretching en la mejora de la calidad de vida y dolor en los pacientes con fibromialgia, Atención Primaria 44 (11) (2012) 641–649.
[20] E.G. d. S. Borges, R.G. d. S. Vale, C.S. Pernambuco, S.A. Cader, S.P.C. Sá, F.M. Pinto, I.C.R. Regazzi, V.M. d. A.O. Knupp, E.H.M. Dantas, Effects of dance on the postural balance, cognition and functional autonomy of older adults, Revista brasileira de enfermagem 71 (2018) 2302–2309.
[21] Y. Zhu, H. Wu, M. Qi, S. Wang, Q. Zhang, L. Zhou, S. Wang, W. Wang, T. Wu, M. Xiao, et al., Effects of a specially designed aerobic dance routine on mild cognitive impairment, Clinical interventions in aging 13 (2018) 1691.
[22] R. Pinniger, R.F. Brown, E.B. Thorsteinsson, P. McKinley, Argentine tango dance compared to mindfulness meditation and a waiting-list control: A randomised trial for treating depression, Complementary therapies in medicine 20 (6) (2012) 377–384.
[23] K.K. Patterson, J.S. Wong, T.-U. Nguyen, D. Brooks, A dance program to improve gait and balance in individuals with chronic stroke: a feasibility study, Topics in Stroke Rehabilitation 25 (6) (2018) 410–416.
[24] S. Hsueh, S.F. Alaoui, W.E. Mackay, Understanding kinaesthetic creativity in dance, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–12.
[25] M. Rüth, K. Kaspar, Exergames in formal school teaching: A pre-post longitudinal field study on the effects of a dance game on motor learning, physical enjoyment, and learning motivation, Entertainment Computing 35 (2020).
[26] A. Romero-Hernandez, M. Gonzalez-Riojo, M. El Yamri, B. Manero, The effectiveness of a video game as an educational tool in incrementing interest in dance among younger generations.
[27] A.D. Kloos, N.E. Fritz, S.K. Kostyk, G.S. Young, D.A. Kegelmeyer, Video game play (dance dance revolution) as a potential exercise therapy in huntington's disease: a controlled clinical trial, Clinical rehabilitation 27 (11) (2013) 972–982.
[28] M. Adcock, F. Sonder, A. Schättin, F. Gennaro, E.D. de Bruin, A usability study of a multicomponent video game-based training for older adults, European review of aging and physical activity 17 (1) (2020) 1–15.
[29] JustDance (video game series). URL: https://en.wikipedia.org/w/index.php?title=Just_Dance_(video_game_series)&oldid=1121254502.
[30] DanceDance Revolution. URL: https://en.wikipedia.org/w/index.php?title=Dance_Dance_Revolution&oldid=1118129088.
[31] DanceCentral. URL: https://en.wikipedia.org/w/index.php?title=Dance_Central&oldid=1116869192.
[32] BeatSaber. URL: https://en.wikipedia.org/w/index.php?title=Beat_Saber&oldid=1120839189.
[33] Synth Riders - A Freestyle-Dance VR Rhythm Game. URL: https://synthridersvr.com/.
[34] DANCE COLLIDER. URL: https://www.dancecollider.com.
[35] S.F. Alaoui, F. Bevilacqua, B.B. Pascual, C. Jacquemin, Dance interaction with physical model visuals based on movement qualities, Int. J. Arts Technol. 6 (2013) 357–387.
[36] S.F. Alaoui, B. Caramiaux, M. Serrano, F. Bevilacqua, Movement qualities as interaction modality, in: DIS '12, 2012.
[37] J. Zhang, Z. Tu, J. Yang, Y. Chen, J. Yuan, Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13232–13242.
[38] W. Li, H. Liu, H. Tang, P. Wang, L. Van Gool, Mhformer: Multi-hypothesis transformer for 3d human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13147–13156.
[39] Y. Xu, J. Zhang, Q. Zhang, D. Tao, Vitpose: Simple vision transformer baselines for human pose estimation (2022). doi:10.48550/ARXIV.2204.12484. URL: https://arxiv.org/abs/2204.12484.
[40] B. Rim, N.-J. Sung, J. Ma, Y.-J. Choi, M. Hong, Real-time human pose estimation using rgb-d images and deep learning, Journal of Internet Computing and Services 21 (3) (2020) 113–121.
[41] V. Srivastav, T. Issenhuth, A. Kadkhodamohammadi, M. de Mathelin, A. Gangi, N. Padoy, Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation, arXiv preprint arXiv:1808.08180 (2018).
[42] D. Michel, A. Qammaz, A.A. Argyros, Markerless 3d human pose estimation and tracking based on rgbd cameras: an experimental evaluation, in: Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, 2017, pp. 115–122.
[43] J. Zhang, Y. Chen, Z. Tu, Uncertainty-aware 3d human pose estimation from monocular video, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5102–5113.
[44] H. Shuai, L. Wu, Q. Liu, Adaptive multi-view and temporal fusing transformer for 3d human pose estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).
[45] H. Tu, C. Wang, W. Zeng, Voxelpose: Towards multi-camera 3d human pose estimation in wild environment, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 197–212.
[46] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, P.V. Fua, Learning monocular 3d human pose estimation from multi-view images, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8437–8446.
[47] L. Ge, H. Liang, J. Yuan, D. Thalmann, Robust 3d hand pose estimation in single depth images: From single-view cnn to multi-view cnns, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3593–3601.
[48] M. Omran, C. Lassner, G. Pons-Moll, P.V. Gehler, B. Schiele, Neural body fitting: Unifying deep learning and model based human pose and shape estimation, in: 2018 International Conference on 3D Vision (3DV), 2018, pp. 484–494.
[49] Y. Li, H. Zhou, B. Yang, Y. Zhang, Z. Cui, H. Bao, G. Zhang, Graph-based asynchronous event processing for rapid object recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 934–943.
[50] J. Kim, J. Bae, G. Park, D. Zhang, Y.M. Kim, N-imagenet: Towards robust, fine-grained object recognition with event cameras, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2146–2156.
[51] I. Alonso, A.C. Murillo, Ev-segnet: Semantic segmentation for event-based cameras, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[52] Ö. Yílmaz, C. Simon-Chane, A. Histace, Evaluation of event-based corner detectors, Journal of Imaging 7 (2) (2021) 25.
[53] S.A. Mohamed, J.N. Yasin, M.-H. Haghbayan, A. Miele, J. Heikkonen, H. Tenhunen, J. Plosila, Dynamic resource-aware corner detection for bio-inspired vision sensors, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 10465–10472.
[54] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, H. Wen, Ev-gait: Event-based robust gait recognition using dynamic vision sensors, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6351–6360.
[55] R. Ghosh, A.K. Gupta, A.N. Silva, A.B. Soares, N.V. Thakor, Spatiotemporal filtering for event-based action recognition, ArXiv abs/1903.07067 (2019).
[56] V. Brebion, J. Moreau, F. Davoine, Real-time optical flow for vehicular perception with low-and high-resolution event cameras, IEEE Transactions on Intelligent Transportation Systems (2021).

[57] M. Liu, T. Delbruck, Edflow: Event driven optical flow camera with keypoint detection and adaptive block matching, IEEE Transactions on Circuits and Systems for Video Technology Epub–ahead (2022).
[58] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, D. Scaramuzza, Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction, IEEE Robotics and Automation Letters 6 (2) (2021) 2822–2829.
[59] J. Hidalgo-Carrió, D. Gehrig, D. Scaramuzza, Learning monocular dense depth from events, in: 2020 International Conference on 3D Vision (3DV), 2020, pp. 534–542.
[60] J. Jiao, H. Huang, L. Li, Z. He, Y. Zhu, M. Liu, Comparing representations in tracking for event camera-based slam, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1369–1376.
[61] J. Bertrand, A. Yiğit, S. Durand, Embedded event-based visual odometry, in: 2020 6th International Conference on Event-Based Control, Communication, and Signal Processing (EBCCSP), IEEE, 2020, pp. 1–8.
[62] J. Li, S. Dong, Z. Yu, Y. Tian, T. Huang, Event-based vision enhanced: A joint detection framework in autonomous driving, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 1396–1401.
[63] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrbein, A. Knoll, Event-based neuromorphic vision for autonomous driving: a paradigm shift for bio-inspired visual sensing and perception, IEEE Signal Processing Magazine 37 (4) (2020) 34–49.
[64] A. Manilii, L. Lucarelli, R. Rosati, L. Romeo, A. Mancini, E. Frontoni, 3d human pose estimation based on multi-input multi-output convolutional neural network and event cameras: A proof of concept on the dhp19 dataset, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I, Springer, 2021, pp. 14–25.
[65] X. Berthelon, G. Chenegros, T. Finateu, S.-H. Ieng, R.B. Benosman, Effects of cooling on the snr and contrast detection of a low-light event-based camera, IEEE Transactions on Biomedical Circuits and Systems 12 (2018) 1467–1474.
[66] A. Amir, B. Taba, D.J. Berg, T. Melano, J.L. McKinstry, C. di Nolfo, T.K. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J.A. Kusnitz, M.V. DeBole, S.K. Esser, T. Delbrück, M. Flickner, D.S. Modha, A low power, fully event-based gesture recognition system, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7388–7397.
[67] Y. Hu, H. Liu, M. Pfeiffer, T. Delbruck, Dvs benchmark datasets for object tracking, action recognition, and object recognition, Frontiers in Neuroscience 10 (2016).
[68] H. Rebecq, T. Horstschaefer, D. Scaramuzza, Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization (2017).
[69] A.I. Maqueda, A. Loquercio, G. Gallego, N. García, 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, 2018, pp. 5419–5427.
[70] R. Benosman, C. Clercq, X. Lagorce, S.-H. Ieng, C. Bartolozzi, Event-based visual flow, IEEE transactions on neural networks and learning systems 25 (2) (2013) 407–417.
[71] A.Z. Zhu, L. Yuan, K. Chaney, K. Daniilidis, Unsupervised event-based learning of optical flow, depth, and egomotion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 989–997.
[72] X. Lagorce, G. Orchard, F. Galluppi, B.E. Shi, R.B. Benosman, Hots: A hierarchy of event-based time-surfaces for pattern recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1346–1359.
[73] R.W. Baldwin, R. Liu, M.B. Almatrafi, V.K. Asari, K. Hirakawa, Time-ordered recent event (tore) volumes for event cameras, ArXiv abs/2103.06108 (2021).
[74] Y. Hu, S.C. Liu, T. Delbruck, v2e: From video frames to realistic DVS events, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2021, URL: http://arxiv.org/abs/2006.07722.
t[75] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[76] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems 28 (2015).
[77] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, 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) (2017) 142–149.
[78] H. Rebecq, D. Gehrig, D. Scaramuzza, Esim: an open event camera simulator, in: Conference on robot learning, PMLR, 2018, pp. 969–982.
[79] D. Joubert, A. Marcireau, N. Ralph, A. Jolley, A. van Schaik, G. Cohen, Event camera simulator improvements via characterized parameters, Frontiers in Neuroscience 910 (2021).
[80] P. Goyal, Q. Duval, I. Seessel, M. Caron, M. Singh, I. Misra, L. Sagun, A. Joulin, P. Bojanowski, Vision models are more robust and fair when pretrained on uncurated images without supervision, 2022.
[81] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, V. Koltun, CARLA: An open urban driving simulator, in: Proceedings of the 1st Annual Conference on Robot Learning, 2017, pp. 1–16.
[82] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, J. Kautz, Super slomo: High quality estimation of multiple intermediate frames for video interpolation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9000–9008.
[83] C. Brandli, R. Berner, M. Yang, S.-C. Liu, T. Delbruck, A 240 180 130 db 3 s latency global shutter spatiotemporal vision sensor, IEEE Journal of Solid-State Circuits 49 (2014) 2333–2341.
[84] W. Falcon, T.P.L. team, Pytorch lightning, the lightweight PyTorch wrapper for high-performance AI research. Scale your models, not the boilerplate. (3 2019). doi:10.5281/zenodo.3828935. URL: https://www.pytorchlightning.ai.
[85] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[86] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, Y.A. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).
[87] H.-J. Yun, K.-I. Kim, J.-H. Lee, H.-Y. Lee, Development of experience dance game using kinect motion capture, KIPS transactions on software and data engineering 3 (1) (2014) 49–56.
[88] M.N. Kamel Boulos, Xbox 360 kinect exergames for health, Games for Health: Research, Development, and Clinical Applications 1 (5) (2012) 326–330.
[89] I. Rallis, A. Langis, I. Georgoulas, A. Voulodimos, N. Doulamis, A. Doulamis, An embodied learning game using kinect and labanotation for analysis and visualization of dance kinesiology, in: 2018 10th international conference on virtual worlds and games for serious applications (VS-Games), IEEE, 2018, pp. 1–8.
[90] A. Kitsikidis, K. Dimitropoulos, S. Douka, N. Grammalidis, Dance analysis using multiple kinect sensors, in: 2014 international conference on computer vision theory and applications (VISAPP), Vol. 2, IEEE, 2014, pp. 789–795.
[91] D.S. Alexiadis, P. Kelly, P. Daras, N.E. O'Connor, T. Boubekeur, M.B. Moussa, Evaluating a dancer's performance using kinect-based skeleton tracking, in: Proceedings of the 19th ACM international conference on Multimedia, 2011, pp. 659–662.
[92] J. Jiao, L. Yuan, W. Tang, Z. Deng, Q. Wu, A post-rectification approach of depth images of kinect v2 for 3d reconstruction of indoor scenes, ISPRS International Journal of Geo-Information 6 (11) (2017), https://doi.org/10.3390/ijgi6110349, https://www.mdpi.com/2220-9964/6/11/349.
[93] Nuitrack Full Body Skeletal Tracking Software. URL: https://nuitrack.com/.
[94] Loihi 2 - Intel - WikiChip. URL: https://en.wikichip.org/wiki/intel/loihi_2.

Zhongyang Zhang received the BS degree in Electronic and Information Engineering from Huazhong University of Science and Technology, Wuhan, Hubei, China, in 2019. She is currently pursuing the PhD degree in Hal?c?o?lu Data Science Institute at UC San Diego. His research interests include the DVS (event camera, neuromorphic camera) imaging, hyperspectral imaging, human pose estimation, and deep learning.

Kaidong Chai is a Master's student in Computer Science at the University of Massachusetts Amherst. He received his Bachelor's degrees in Computer Science and Mathematics from the University of Massachusetts Amherst in 2022. He has engaged in research related to massive data generation and processing, human pose estimation, and educational technology. His research interests include computer vision, machine learning, and operating systems.

Haowen Yu received the MSCS degree from the University of Massachusetts Amherst, specializing in computer vision research. With a particular focus on unsupervised learning, human pose estimation, and reinforcement learning, he has conducted cutting-edge research and developed innovative algorithms in these fields. After completing his studies, Haowen Yu joined the TikTok Global Search Ads Group as a software engineer. His main responsibility is to work on ads mixed ranking, a crucial aspect of TikTok's advertising platform that leverages his expertise in computer vision and reinforcement learning to improve ad targeting and relevance.
Throughout his academic and professional career, Haowen Yu has demonstrated a passion for pushing the boundaries of computer vision and machine learning. With a solid foundation in both theoretical and practical aspects of the field, he is constantly seeking new challenges and opportunities to apply his skills in innovative ways.

Ramzi Majaj received the M.S degree from the University of Memphis, in 2018 with a focus of human biomechanics. He is currently a staff scientist at the Center for Human Health & Performance at the Institute for Applied Life Science. His Research interest are in motion capture biomechanics, exercise physiology, physical activity and health measurements.

Francesca Walsh is a Neuroscience and Behavior Ph.D. Candidate at the University of Massachusetts Amherst. In 2018, she received her B.S. in an Individual Concentration in 'Neurobiology, Economic Decision-making, and Social Systems' with a second major in Economics. In 2019, she received her M.S. in Neuroscience and Behavior from the University of Massachusetts Amherst. She studies decision-making, specifically focusing on economic and subjective value choices. During the COIVD-19 pandemic, she began working in the BINDs lab under Dr. Hava Siegelmann to leverage the brain's decision-making and planning mechanisms in artificial intelligence designs. She has three posters and one publication.

Edward J. Wang, Ph.D., is an Assistant Professor of Design and Electrical and Computer Engineering at UC San Diego. He received his Ph.D. in Electrical and Computer Engineering from the University of Washington in 2019. Dr. Wang is the director of the UC San Diego DigiHealth Lab where he develops novel health monitoring and digital intervention solutions to deliver healthcare in non-traditional settings. He has authored 20+ articles at peer-reviewed journals and conferences, including best paper awards from ACM Ubicomp, CHI, and IEEE Pervasive Health.

Upal Mahbub, Ph.D., is a Senior Member of IEEE and a Staff Engineer at Qualcomm Technologies Inc. He received an M.Sc. and his Ph.D. in Electrical and Computer Engineering from the University of Maryland College Park in 2018 and 2018, respectively. Dr. Mahbub has authored 30+ articles at international conferences and prestigious journals, edited a book titled "Contactless Human Activity Analysis" from Springer, in 2021, and received several awards including best paper awards from IEEE UEMCON 2016 and ICCIT 2011. He is currently developing hardware-efficient CV/ML solutions for different perception tasks in the XR domain.

Ph.D., Computer Science, Rutgers University (1993, Fellow of excellence), M.Sc., Computer Science, Hebrew University (1992, Cum Laude), B.A., Computer Science, the Technion (1988, Suma Cum Laude). Siegelmann has been a visiting professor at MIT, Harvard University, the Weizmann Institute, ETH, the Salk Institute, Mathematical Science Research Institute Berkeley, and the Newton Institute of Cambridge University.
Professor Siegelmann recently completed a four-year term as a PM of some of DARPA's most significant and innovative AI programs: Lifelong Learning Machines "L2M," one of her key initiatives, inaugurated "third-wave AI," pushing major design innovation, inspired by biology, and a dramatic increase in AI capability. "GARD" is leading to novel advancements in assuring AI robustness against attack. "CSL" is introducing powerful methods of collaborative information sharing on AI platforms without revealing private data. Other programs include advanced biomedical applications. DARPA/DoD bestowed upon her the Meritorious Public Service Medal, one of the highest medals for civilians, for her research and leadership.

Donghyun Kim, Ph.D., is an Assistant Professor in the College of Information and Computer Sciences at the University of Massachusetts Amherst. He earned his Ph. D. in Mechanical Engineering from the University of Texas at Austin and now heads the Dynamic and Autonomous Robotics Lab. Dr. Kim's research focuses on enhancing the physical capabilities of legged robots by improving their perception, control, and hardware. He has authored over 25 papers published in top-tier peer-reviewed journals and conferences, including a best paper award from Transactions on Mechatronics.

Tauhidur Rahman is an Assistant Professor in the Hal? c?o?lu Data Science Institute and Computer Science and Engineering at the University of California San Diego where he directs the Mobile Sensing and Ubiquitous Computing Laboratory (MOSAIC Lab). Tauhidur received his Ph.D. degree from Cornell University and his research focuses on building novel ubiquitous and mobile health sensing technologies. Some of his notable accomplishments include a Google Ph.D. fellowship in 2016 in mobile computing, Outstanding Teaching Award 2015 from Cornell University, one best paper awards in ACM Digital Health 2016, ACM Ubicomp 2015 and ACM IMWUT in 2021.