EventPoseFormer: Event-Based Real-time 3D Human Pose Estimation

Melica Omer Ali*
s6021281@studenti.unige.it

Arren Glover†
arren.glover@iit.it

Chiara Bartolozzi†
chiara.bartolozzi@iit.it

Francesca Odone*
francesca.odone@unige.it

Gaurvi Goyal†§
gaurvi.goyal@maastrichtuniversity.nl

***Abstract*—Human Pose Estimation (HPE) provides a structured representation of human motion and has become a key component in a wide range of applications, spanning from medical monitoring and rehabilitation to sports biomechanics and consumer-oriented applications such as dance and yoga instruction. While 2D pose estimation is sufficient for some of these use cases, others require 3D pose information to be effective. Despite significant recent progress in HPE, achieving real-time 3D estimation with low latency is still a challenging problem. This challenge may be addressed through the use of *event cameras* — novel, bio-inspired, visual sensors which react to changes in the intensity values in the scene, asynchronously, at pixel level. This sensing modality is extremely efficient, as it removes redundancy at the sensor level, and is well suited for the design of high-speed, robust artificial vision algorithms, including HPE. In this paper, we combine MoveEnet, a HPE system based on event cameras which performs real-time high-frequency 2D pose estimation, with PoseFormerV2, a transformer-based 2D-to-3D pose lifting model to develop an end-to-end system for monocular real-time 3D HPE. Our method not only performs more accurately and faster than existing HPE models for event cameras, but is also competitive with RGB models while being substantially faster. Our final system has an end-to-end latency of 30ms, a real-time frequency of 38Hz and a 3D MPJPE of 85.7mm on event-Human 3.6M dataset. Code for the end-to-end real-time pipeline can be found at https://github.com/event-driven-robotics/HPE-event-poseformer.**

***Index Terms*—Human pose estimation, event camera, 2D-to-3D lifting, real-time**

I. Introduction

3D Human Pose Estimation (HPE) had been widely researched and incorporated in a variety of applications in the last decade [1]–[4]. Accurate estimation of a person's location within a scene, as well as their body pose and motion over time, is a key source of information for many human-centered applications, including smart homes, autonomous driving, medical diagnosis, rehabilitation, and robotics. Therefore, more works, generalizing to a higher variety in body types, ages, point of views, angles, number of people in the view, etc. have been proposed — see, for instance, the recent survey [1]. While robust and accurate marker-less solutions have been proposed, they are often computationally heavy and therefore unsuitable for small, portable, or battery-constrained devices, and also have a low-frequency output with noticeable lag, making them unsuitable for applications which require real-time feedback, e.g., augmented reality or robot control [5].

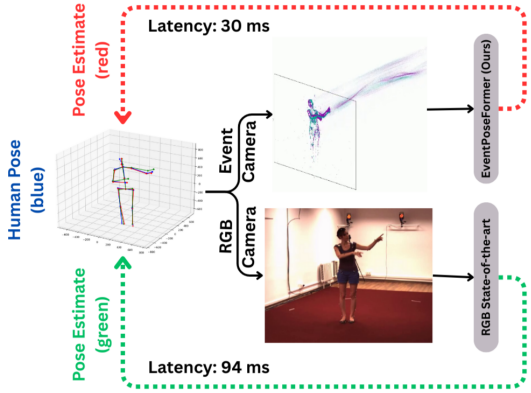


Fig. 1: EventPoseFormer (top red) converts a single event-camera into a 3D Human Pose with 30 ms latency, >3x faster than RGB camera state-of-the-art (bottom green). EventPose-Former can enable real-time feedback for augmented reality and physiological disease diagnosis.

This work focuses on real-time estimation of human pose in 3D, leveraging the low latency, high temporal resolution, and data compression of event cameras, which respond only to *change* in the scene intensity at the pixel level. Event cameras output sparse data streams, or events, by independently detecting changes in brightness at each pixel and asynchronously sending events as soon as the changes are detected. The functionality is fundamentally different from frame-based sensors, where light is integrated for a fixed amount of time and the measured values of all pixels are transmitted together in a large packet. While the event-driven acquisition results in incredibly low latency, high dynamic range, low power, and no motion blur, the latter results in fixed latency, redundant data to process, and motion blur in dynamic scenes, especially in low light conditions. Their properties make them ideal sensors for real-time applications for long term and continual monitoring of dynamical scenes, where motion blur affects mainstream vision sensors, in a wide variety of illumination conditions, such as in most applications of (3D) HPE. Yet, 3D HPE with

*DIBRIS, University of Genova, Italy
†Event Driven Perception for Robotics Istituto Italiano di Technologia, Italy
§Department of Computing Sciences Maastricht University, Netherlands

event cameras is scarcely explored at this time (e.g., [6], [7]).

In this paper, we propose an end-to-end pipeline that views a human subject with a single event camera, and estimates their 3D pose in real-time, see Fig. 1. A model for event-based real-time 2D HPE, MoveEnet [8], is paired with a 2D-to-3D lifting model inspired by PoseFormerV2 [9]. As the two models use different joint model specifications, a joint extrapolation module is introduced (PoseBridge), to transform the output of MoveEnet into a compatible input for PoseFormerV2. Additionally, PoseFormerV2 depends on both past and future frames to predict 3D poses. In our approach, we present a variation of this model that eliminates the need for having future data, thus making it suitable for real-time low-latency processing.

In summary, the contributions of this paper are:

- An integrated real-time pipeline for 3D human pose estimation from a single event camera.
- A novel PoseBridge module for transforming 13-joint 2D poses from event-based estimation into a 17-joint 2D poses for 3D pose lifting.
- A 2D-to-3D lifting module designed for real-time, low-latency operation that only relies on present and past input data.
- A full evaluation of the accuracy of the proposed event camera based 3D HPE system.

II. Related Work

Recent research has explored a wide range of approaches for reconstructing 3D human pose from images. Existing methods can be broadly classified into end-to-end models that directly estimate 3D poses from images [10]–[12] and two-stage pipelines that decouple 2D pose estimation from subsequent 3D pose lifting [4], [13], [14]. Despite progress in end-to-end learning, 2D-to-3D lifting remains attractive due to its modularity, robustness, and adaptability to different sensing modalities.

3D Human Pose Estimation from RGB input Among the two-stage approaches, it is worth mentioning the ones leveraging multiple views: multiple 2D HPE can be computed on the first stage, while in the second stage, 3D HPE can be computed with geometrical reasoning or machine learning. For instance, in [15], the authors propose a markerless gait analysis framework that reconstructs 3D joint trajectories from multi-view RGB videos by first estimating 2D keypoints and then applies geometric triangulation. Similarly, AdaFuse fuses 2D HPE from multiple views with a joint-learning geometric approach [16]. The Multi-view Pose Lifter (MPL) [17] proposed a learning-based multi-view lifting method to overcome some limitations of geometric reconstruction and employed transformer-based architectures to fuse 2D poses from multiple calibrated views into a unified 3D representation. While such approaches achieve high accuracy and allow precise joint estimation, they require multiple calibrated cameras and are therefore not well suited for monocular scenarios.

Monocular 3D HPE has been fueled by recent advances in deep learning, as summarized in [1]. Despite the feasibility of the task, the accuracy and applicability of the obtained results is still unclear. Several studies [1], [3], [18]–[21] have investigated the influence of 2D pose quality on downstream 3D reconstruction performance. In [19], recent state-of-the-art 2D pose estimators are evaluated in combination with state-of-the-art 2D-to-3D lifting models. The results demonstrate that improvements in 3D accuracy are strongly correlated with the reliability of 2D keypoints, and that noise reduction at the 2D level can significantly enhance lifting performance. Transformer-based lifters such as PoseFormerV2 [9] and graph-based models such as MotionAGFormer [22] effectively capture spatial and temporal dependencies in pose sequences.

Overall, the majority of existing work on 3D HPE relies on conventional frame-based RGB cameras. In contrast, event-based vision has only recently begun to receive attention in this context, hence, event-based methods for HPE remain comparatively limited.

2D Human Pose Estimation from events Among event-based approaches, MoveEnet [8] is the state-of-the-art method for real-time 2D HPE using an event camera. The method adapts a convolutional neural network (CNN) architecture to operate on event representations, enabling high-frequency pose estimation. MoveEnet demonstrates that event-driven sensing can achieve competitive 2D pose accuracy with the appropriate representation and training paradigms. GraphEnet [23] presents a Graph Neural Network that takes line segments fitted to the event data as the input, then builds a graph with the line segments and processes it with a GNN to estimate the 2D human pose. The topic of 2D HPE is minimally explored at this time, with few works available. While GraphEnet has lower latency, MoveEnet is more accurate in its results.

3D Human Pose Estimation from events The earliest work, [24] estimates 2D pose by supplying event histograms to CNNs to obtain 2D pose estimation for each camera individually, then geometrically triangulating the poses to estimate 3D pose. While being an effective technique, this requires multiple, synchronized cameras for input. EventHPE [25] proposes a two-stage pipeline that first estimates optical flow from events using an unsupervised network and then predicts 3D human pose and shape from both the event representation and the inferred flow. A monocular approach is presented in [6], which introduces a learning-based framework for estimating 3D human pose by predicting orthogonal heatmaps for each joint and subsequently fusing them to recover 3D joint locations. EvTransPose [26] presents a transformer based approach for the task with multiple hourglass modules placed sequentially in the pipeline, and focusing on the spatial relationship of the joints. Although all three of these methods are effective at the task, they are computationally expensive with long times to process individual inputs, and therefore not applicable to a real-time scenario.

[27] instead, exploits the temporal resolution and sparsity of events, with a 3 channel hybrid spiking neural network(SNN) approach to extract spatial, temporal and event-based features, to estimate human pose. SNNs are highly efficient when deployed on appropriate, custom hardware for Neuromorphic

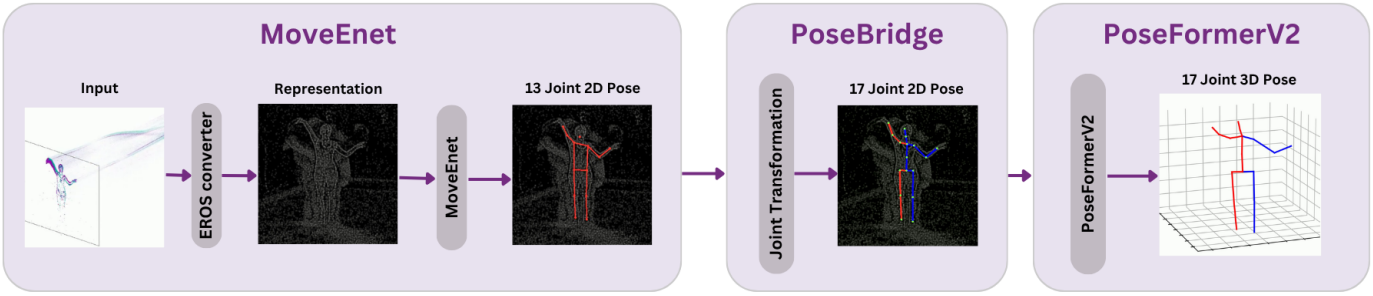


Fig. 2: Full Pipeline of EventPoseFormer. MoveEnet [8] is used to produce low-latency (11 ms) 2D joint positions from event camera data. PoseBridge converts the 13 Joint model to a 17 Joint model, and PoseFormerV2 is a modified lifting network that focuses on maintaining low-latency and real-time performance. The final result is 17 joints in 3D Cartesian space from a single event camera.

computing, but lose their efficiency on conventional von Neumann hardware.

Other works have also been introduced, focusing on adjacent but related problems with event cameras. For example, [7], focused on egocentric HPE, that is, estimating the pose of a subject from data acquired from a head-mounted camera worn by the subject themselves. This work emphasizes the advantages of event-driven sensing, particularly in terms of real-time processing and robustness to motion blur, even with a moving camera. In case of occlusion, this method comes up with best estimates for the occluded joints. EventHands [28] estimates the 3D pose mesh of a human hand with event camera data at a very high frequency in real time. Another work [29] incorporates event camera data to generate an RGB dataset with motion blur, in order to augment the training, thus promoting robustness to motion blur in RGB images for 2D pose estimation.

In this work, MoveEnet [8] is adopted as the event-based 2D pose estimation backbone, due to the high accuracy which is crucial to support the downward module. By coupling its high-frequency 2D pose output with transformer-based 2D-to-3D pose lifting models, the proposed system extends event-based HPE into the 3D domain while preserving a modular and real-time pipeline.

III. Methodology

We design a multi-stage pipeline, depicted in Fig. 2, consisting of the event-camera input, the 2D joint estimation with MoveEnet, the conversion of the 13 joint model to the 17 joint model, and the lifting of the 2D joint positions to 3D.

A. Event Camera

The event camera produces an event stream $\mathcal{V} = \{e_i\}_{i=1}^{N}$, where each event $e_i = (x_i, y_i, t_i, p_i)$ consists of pixel coordinates (x_i, y_i), timestamp t_i, and polarity p_i. Events occur at microsecond temporal resolution enabling millions of finite change measurements to occur each second.

B. 2D pose estimation: MoveEnet

To perform the initial 2D estimation of joint positions of a single human, we process the event stream with MoveEnet [8]: from the variable amount of event data, MoveEnet first computes EROS (Exponentially-Reduced Ordinal Surface) [30], a matrix representation of the event data. EROS serves three purposes, 1. keeping memory of joints that have not moved since the last update, 2. removing integration blur for fast moving joints, and 3. providing the constant size input data required by neural networks.

EROS is invariant to the motion of the person: fast moving body parts (e.g. hands) are represented equally sharp as other body parts that do not move, or move slower (e.g. trunk, or legs when the person is not walking). The 2D EROS representation, $\mathcal{E} : (x, y) \mapsto [0.0, 1.0]$, at any timestep, t_i, is computed from all events in the event stream from $t = 0$:

$$\mathcal{E}_{t_i} = f_{\text{EROS}}(\mathcal{V}_{t=0}^{t_i})$$

where $f_{\text{EROS}}(.)$ is defined in [30]. $\mathcal{E}_{t_i}$ can be sampled for any value of t_i equal to the event stream temporal resolution. Additionally $\mathcal{E}$ is computed incrementally for new data e_{i+1}.

The MoveEnet joint position estimation network, W_{MoveEnet}, is an hourglass shaped CNN with multiple regression heads, creating heatmaps that are combined to obtain 13 body joints, $\{J\}^{\mathbf{13}}$, i.e.

$$\{J\}_{t_i}^{\mathbf{13}} \leftarrow W_{\text{MoveEnet}}(\mathcal{E}_{t_i})$$

MoveEnet was chosen as it can run at a frequency of up to 100 Hz using a GPU and therefore takes advantage of the event camera and the high sampling rate of $\mathcal{E}$.

C. Extrapolation module: PoseBridge

MoveEnet produces 2D poses with a 13-joint skeleton, $\{J\}^{\mathbf{13}}$, which must be extrapolated to the 17-joint format, $\{J\}^{\mathbf{17}}$, required by the 3D pose lifting module using the proposed *PoseBridge*, i.e.

$$\{J\}_{t_i}^{\mathbf{17}} = f_{\text{PoseBridge}}(\{J\}_{t_i}^{\mathbf{13}})$$

In this mapping, the *head* joint provided by MoveEnet corresponds to the *nose* joint in the 17-joint skeleton. The joints not predicted by MoveEnet, namely the neck, head, hip and spine, are reconstructed with the linear combinations of

other predicted joints as outlined in Equations 1, 2, 3 and 4, respectively:

$$J^{\mathbf{17}}_{\text{neck}} = \frac{J^{\mathbf{13}}_{\text{LShoulder}} + J^{\mathbf{13}}_{\text{RShoulder}}}{2} \quad (1)$$

$$J^{\mathbf{17}}_{\text{head}} = 2 \cdot J^{\mathbf{13}}_{\text{nose}} - J^{\mathbf{17}}_{\text{neck}} \quad (2)$$

$$J^{\mathbf{17}}_{\text{hip}} = \frac{J^{\mathbf{13}}_{\text{RHip}} + J^{\mathbf{13}}_{\text{LHip}}}{2} \quad (3)$$

$$J^{\mathbf{17}}_{\text{spine}} = \frac{J^{\mathbf{17}}_{\text{neck}} + J^{\mathbf{17}}_{\text{hip}}}{2} \quad (4)$$

D. 2D-to-3D Lifting: PoseFormerV2

The 3D reconstruction stage is modeled as a 2D-to-3D pose lifting problem, built upon PoseFormerV2 [9]. Given a temporal sequence of 2D joint coordinates from time t_j to t_i, $\{J\}^{\mathbf{17}}_{t_j \to t_i}$, we first normalize the 2D joint poses in image space, producing $\{\bar{J}\}^{\mathbf{17}}_{t_j \to t_i}$ according to:

$$x_{\text{norm}} = \frac{x_{\text{raw}}}{w} \times 2 - 1 \quad (5)$$

$$y_{\text{norm}} = \frac{y_{\text{raw}}}{w} \times 2 - \frac{h}{w} \quad (6)$$

where w and h are the width and height of the image.

PoseFormerV2 estimates the corresponding 3D joint coordinates, $\{X\}^{\mathbf{17}}_{t_i}$, by modeling spatial relationships between body joints and temporal motion patterns across frames

$$X^{\mathbf{17}}_{t_i} = W_{\text{PoseFormerV2}}(\{\bar{J}\}^{\mathbf{17}}_{t_j \to t_i})$$

giving the final joint poses in 3D cartesian space. The model adopts a transformer-based architecture, where self-attention mechanisms allow for efficient modeling of long-range temporal dependencies in the pose sequence.

To make our module compatible with real-time inference, a constraint was placed that the model should only know past information. The original PoseFormerV2 used $M = i-j$ input poses such that $j < t < i$, i.e. poses that existed in the future of the estimated frame at time t were used. In our model, we set all M frames as current and past frames only, i.e. $t = i$ and $j < i$. We propose this as a **causal** PoseFormerV2.

Since the model relies solely on pose information rather than raw image data, it is independent of the modality used to generate the 2D pose estimates. This property makes PoseFormerV2 well-suited for 3D lifting the event-based 2D pose estimation from MoveEnet.

Implementation details: All components of the 3D HPE pipeline are implemented using PyTorch, and inference is performed sequentially, which allows independent evaluation of the 2D and 3D stages. All training parameters are directly adopted from the original papers of MoveEnet [8] and PoseFormerV2 [9]. The event camera used is Prophesee evaluation kit EVK1 with an array size of 640x480, pixel pitch of 15 µm, optical format of 3/4", typical latency of 200 µs and events temporal resolution of 1 µs. The processing is carried out on a Dell Alienware m15 R3, Intel Core i9-10980HK @ 2.40GHz x 16, NVIDIA GeForce RTX 2070, 8 GB DDR4 @ 2667MHz. Causal pose input was achieved by padding all future frames with the current frame, to keep a constant input size.

IV. Experiments

A. Dataset

Event-Human 3.6 Million: The Event-Human3.6M dataset (eH36m) [31] is a semi-synthetic dataset created by converting the original RGB Human3.6M dataset [32] to events with the V2E model [33]. The Human3.6M dataset (H36m) is a widely used large scale, multi-view, benchmark video dataset for single person 2D and 3D HPE, and enables benchmarking against RGB-based methods. The event dataset has a resolution of 640x480 pixels, with 7 subjects (5 for training and 2 for testing) recorded doing 17 different actions, and uses 2 of the 4 camera views of the original dataset.

Warm-up: Error calculations do not consider the first 0.2 seconds of each sequence in the eH36m dataset. Each sample of the original RGB dataset is initialized with the subject standing motionless in the t-pose. As no movement occurs in this time frame, the event-based camera does not produce data, the estimation cannot be performed and evaluation is invalid. As soon as the subject moves, the pose estimation is possible. Such an event camera warm-up is a problem only for recorded datasets and would not exist in real-world conditions, where the event stream is continuous over a long period of time. Thus, removal of this time period from the analysis does not affect the interpretation and validity of the results for real time applications. This time is discarded and not considered in any downstream component including the past frames of the 2d-to-3D lifting module.

V. Evaluation Metrics

A. Metrics

a) Mean Per Joint Position Error (MPJPE): is the measure of the average Euclidean distance between the predicted and ground truth joint positions. It is the primary metric for evaluating 3D pose estimation accuracy.

b) Percentage of Correct Keypoints (PCK): presents the percentage of correct joint localizations. A predicted joint is considered correct if its Euclidean distance from the ground truth position falls below a predefined threshold. In this work, results are reported at a threshold of 150 mm.

c) Frequency: Frequency is measured as the maximum number of poses estimated by a system per second. Some components of the pipeline can be implemented to run in parallel, boosting frequency.

d) Latency: System latency is measured for the whole pipeline. Latency is defined as the time between a change occurring in the world and the system processing and giving the 3D pose as an output. Therefore, latency includes the sensor latency as well as the algorithm latency. A comparison of sensor latency between RGB and event-cameras was presented in [34] with 36.5 ms and 3.5 ms respectively. The latencies of all methods reported in this paper are measured on the same hardware (detailed in implementation details) under similar processing conditions.

Method	Input	MPJPE (mm)		Freq (Hz)	Latency (ms)
		Full	Lifting		
Cheng et al. [35]	RGB	40.1	–	–	–
HRNet-PoseFormerV2 [9]	RGB	69.17	45.2	17.3	94
HRNet-MotionAGFormer [22]	RGB	–	38.4	9.8	138
Scarpellini et al. [6]	Events	116.4	–	2	503
EventPoseFormer **(ours)**	Events	85.7	45.2	38	30

TABLE I: Quantitative comparison of EventPoseFormer with other event and RGB based methods on Human3.6M. Full MPJPE refers to MPJPE of the full pipeline, with events as input and Lifting MPJPE is the reported MPJPE of the lifting model only, with ground truth 2D pose as input. Latency includes contributions from both the sensor and the algorithm.

B. Comparison with state-of-the-art models

We compared the performance of our modular architecture to the methods with available results on event-Human 3.6M. Scarpellini et al. [6], being an end-to-end approach provides insights into the trade-offs of modular versus end-to-end event-based systems. Since the original dataset Human 3.6M is an RGB video dataset, results for the 3D HPE task on this dataset are also presented to enhance the comparison. Results are shown in Tab. I. All MPJPE values are based on scores presented in the original papers and have not been replicated by the authors of this work.

Although RGB-based methods achieve notably higher accuracy in 3D human pose estimation, as summarized in Tab. I, they operate at significantly lower rates. For example, the 2D detection pipelines underlying these 3D algorithms, MoveEnet and HRNet, run at 100 Hz and 20 Hz respectively. The benefit is not as great when moving to 3D, with 38 Hz and 17 Hz respectively, but still significant. The key advantage of a faster sensor and algorithm occurs when considering real-time systems where the pose is used in close-loop with the user (e.g. augmented reality). For example, in Fig. 3, it can be seen that despite the lower accuracy, EventPoseFormer achieves a closer match to ground truth compared with HRNet-PoseFormerV2 when latency is considered in the inaccuracy of a method. Similarly, the quantitative comparison presented in Tab. II shows that when latency is incorporated into the evaluation, the performance gap between the two approaches decreases. This demonstrates that while RGB-based methods may achieve slightly higher spatial accuracy, event-based systems offer significant advantages in responsiveness and temporal alignment with the observed motion.

C. Real-time Inference

The online experimental setup involves a subject standing in front of the event camera, where the system continuously processes incoming event data to estimate the subject 3D pose. Sensor latency, that is, time taken from the stimulus to the beginning of process pipeline is 3.5 ms [34]. Time taken to process the events to the EROS frame is negligible (0.1 μs per event) compared to the other components in this pipeline.

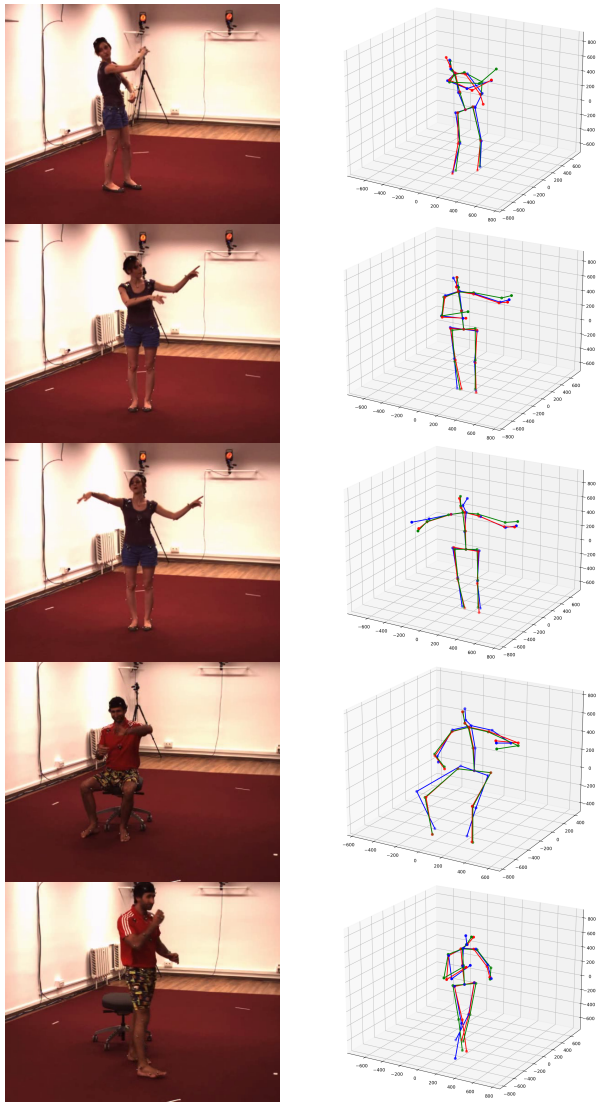


Fig. 3: Latency comparison between EventPoseFormer (ours) (red) and HRNet-PoseFormerV2 (green) shown against ground-truth (blue). In a real-time scenario, EventPoseFormer outputs a pose temporally closer to the ground-truth pose. In contrast, HRNet-PoseFormerV2 gives an estimate later, creating a potentially noticeable lag for the user.

Method	Input	MPJPE (mm)	PCK (%)	MPJPE w/ latency (mm)	PCK w/ latency (%)
HRNet-PoseFormerV2 [9]	RGB	69.17	92.17	74.10	90.40
EventPoseFormer **(ours)**	Events	85.7	85.99	86.58	85.64

TABLE II: Quantitative latency-based MPJPE and PCK@150mm comparison of EventPoseFormer with HRNet-PoseFormerV2 on Human3.6M.

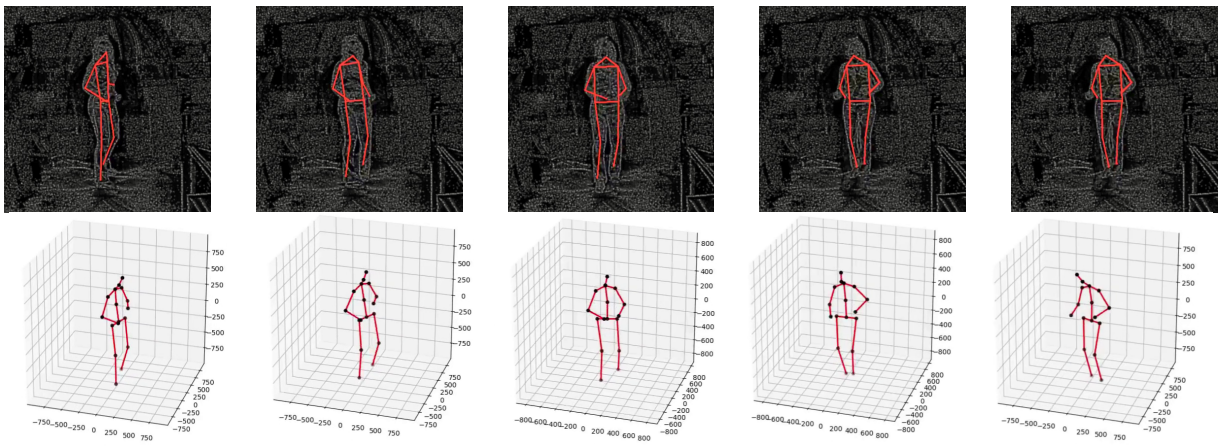


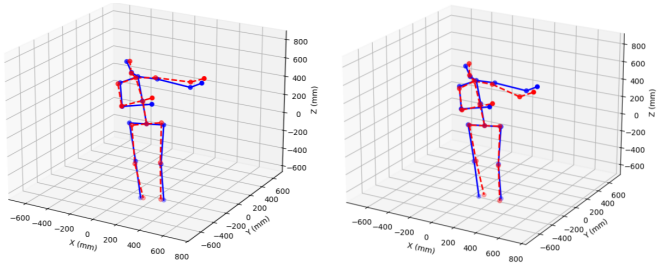
Fig. 4: Snapshots from a real-time demo of EventPoseFormer (originally recorded from a screen-capture). 2D skeleton from MoveEnet superimposed on EROS representation (top) and 3D reconstruction (bottom).

For the online implementation, MoveEnet processes a single EROS event frame in 11 ms, while the subsequent PoseBridge, normalization step and the modified PoseFormerV2 model inference require an additional 15 ms to produce the corresponding 3D pose. Therefore the algorithm latency is measured as 26 ms (38 Hz). The full pipeline latency, including the event camera latency is 30 ms. The proposed system is capable of real-time and online 3D human pose estimation with low latency using a single event camera, as demonstrated in Fig 4.

D. Component-wise ablation

a) 2D backbone: MoveEnet has high accuracy for 2D HPE while retaining relatively low latency, making it an ideal candidate for this work. Nevertheless, to evaluate the performance and suitability of MoveEnet as a 2D backbone for 3D pose estimation, 2D poses generated by MoveEnet from event data are compared against 2D poses produced from RGB input by HRNet [36], a widely adopted state-of-the-art RGB-based pose estimator. For the same sample from eH36m and H36m respectively, 2D poses estimated by MoveEnet and HRNet are provided as input to PoseFormerV2. An instance of the results is shown in Fig. 5. The estimation on a test sample with MoveEnet for the 2D pose estimation resulted in a slightly higher, yet still competitive, MPJPE of 70 mm, in comparison to using HRNet as the 2D estimator, which achieved an MPJPE of 60 mm. Since the data has been converted from video to events, and such a conversion cannot capture the true high-frequency information, we consider the RGB converted error an upper-bound; true event-based data should have increased performance still. Nevertheless, this result indicates that MoveEnet provides sufficiently accurate 2D pose estimates to serve as a reliable input for the proposed 3D pose estimation pipeline. This observation is in line with the results presented in the original paper [8].

b) 2D-to-3D model: For the 3D reconstruction stage, we compared two state-of-the-art transformer-based lifting models, PoseFormerV2 and MotionAGFormer [22]. These models are chosen because of their strong performance in recent benchmarks and because they are designed to operate directly on temporal sequences of 2D joint coordinates, which makes them compatible with event-based pose output and our modular approach. While MotionAGFormer incorporates graph-based reasoning in addition to transformer layers, PoseFormerV2 adopts a simpler transformer architecture with frequency-domain modeling.



(a) HRNet–PoseFormerV2 (b) MoveEnet–PoseFormerV2

Fig. 5: Qualitative comparison of 3D pose estimation pipelines using frame-based and event-based 2D backbones combined with PoseFormerV2. Predicted poses (red) and ground truth (blue).

Both models are evaluated using identical 2D pose inputs generated by HRNet, for a fair comparison. Results are incorporated into Tab. I. Although MotionAGformer is more accurate, it has a substantially higher latency. Considering that the primary goal of this work was to create a real-time implementation, PoseFormerV2 was selected for the final pipeline.

c) Causal and non-causal inference: In the causal setting [37], only past frames are used to estimate the current 3D pose, enabling real-time 3D HPE application. In the non-causal setting, both past and future frames are used to estimate the 3D pose of the current frame. This comparison presents the impact of temporal context on accuracy.

Tab. III presents results of this comparison on two randomly selected samples from the validation set of the eH36m dataset. The results show that non-causal inference generally performs slightly better than causal inference, likely due to the avail-

Subject-Action	Metric	None	Warm-up	Causal+ Warm-up
S9-Eating	MPJPE (mm)	70.98	**64.73**	65.74
	PCK (%)	93.33	**95.84**	95.02
S11-Greeting	MPJPE (mm)	87.97	**79.67**	81.24
	PCK (%)	83.86	**87.64**	87.16

TABLE III: Combined MPJPE and PCK@150mm results for different ablation settings: *All* (using all frames), *Warm-up* (first 0.2s data skipped), and *Causal-warmup* (first 0.2s data skipped with past context only). Best results are in **bold**.

ability of future temporal context. However, the performance gap is relatively small, indicating that the proposed pipeline remains effective in real-time applications where future frames are absent.

d) Warm-up: As explained in Section IV-A, in the first 0.2 seconds of the dataset, the events are sparse and not informative of the pose. Therefore, to improve the reconstruction accuracy of the final 3D output, the first 0.2s of data are skipped of all test sequences to allow the event camera sufficient time to capture the full body motion. As can be seen from the results in Tab. III, across both evaluated actions and subjects, skipping the initial data consistently leads to improved performance. Thus, at steady state, our method performs robustly, supporting its utility in real-time applications.

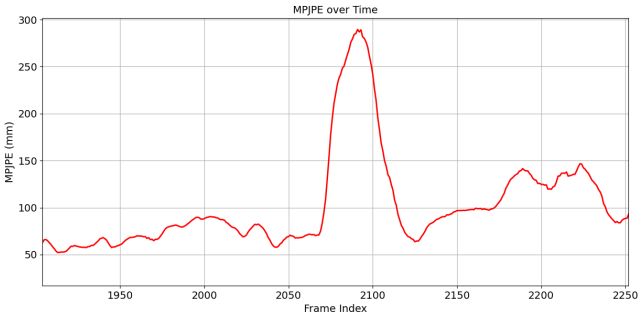
E. Limitations

As shown in the MPJPE over time plot in Fig. 6a, an anomalous spike appears at specific time instants. A closer inspection of the corresponding frames reveals that these spikes occur when the subject rotates sideways or turns their back toward the camera. In these frames, the estimated 3D pose tends to incorrectly assume a frontal body orientation. This behavior is clearly illustrated in the qualitative visualizations, Fig. 6b and Fig. 6c, where the right side of the skeleton (colored in red) and the left side (colored in blue) become swapped in the back-facing pose, likely due to the EROS frame, which lacks the detail required by the 2D HPE differentiate between front view and back view of a person.

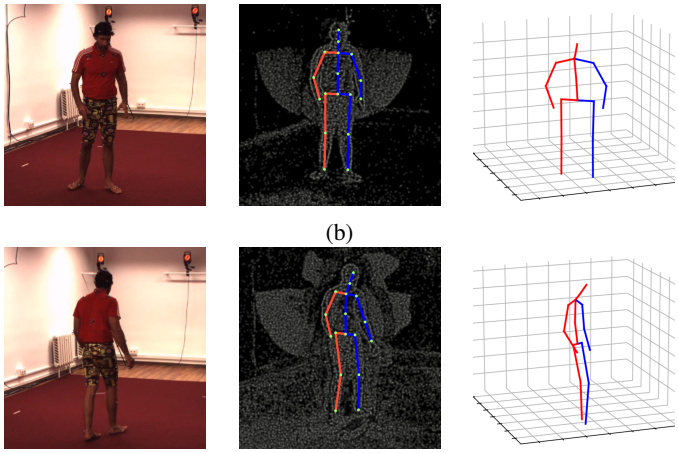
Side-view poses also exhibit reduced accuracy. This is primarily caused by self-occlusion, where one side of the body becomes partially or fully invisible to the event camera. These errors propagate from the 2D pose estimation through the 2D-to-3D lifting stage, leading to less reliable 3D reconstructions.

VI. Conclusion and Future Work

In this work, we presented a real-time pipeline for event-based 3D human pose estimation by combining the strengths of high-frequency event sensing and transformer-based temporal modeling, with MoveEnet for 2D pose estimation and PoseFormerV2 for 2D-to-3D lifting. To bridge the structural mismatch between event-based 2D pose estimation and 3D pose lifting modules, we proposed PoseBridge, a lightweight joint extrapolation module that transforms 13-joint 2D poses



(a)



(b)

(c)

Fig. 6: The (a) MPJPE of the S9-Directions sample with an anomalous spike in error which was attributed to (b) The correct estimation of a frontal view from Frame 121, versus the (b) incorrect estimation of a back-facing view from Frame 2075. The blue and red sides of the skeleton should invert when facing backwards.

into a 17-joint format suitable for 3D reconstruction. In addition, we implemented a causal variant of PoseFormerV2, which excludes future frame information and enables real-time, low-latency inference.

Through quantitative and qualitative evaluations, we showed that the proposed system achieves competitive 3D reconstruction accuracy. Although RGB-based approaches still outperform event-based methods in terms of absolute accuracy, our results highlight a favorable trade-off between accuracy and speed, which is critical for applications requiring fast perception and reaction. Nevertheless, several challenges remain. Performance degradation is observed in side-view and back-facing poses due to occlusion and orientation ambiguity inherent to single camera setups. Future work will focus on addressing these limitations through improved orientation modeling and the exploration of multi-view or multi-person human pose estimation.

Overall, this work demonstrates that event-based sensing is a viable and promising option for real-time 3D human pose estimation, and adopting a modular approach preserves the strengths of each component while allowing independent

training and optimization, paving the way for scalability and future extensions.

ACKNOWLEDGMENT

The authors acknowledge the support of the European Union Horizon Europe programme, grant agreement ID: 101120727, "PRIMI: Performance in Robots Interaction via Mental Imagery" and the support of Air Force Office of Scientific Research, Project number FA8655-23-1-7074.

REFERENCES

[1] Y. Guo, T. Gao, A. Dong, X. Jiang, Z. Zhu, and F. Wang, "A survey of the state of the art in monocular 3d human pose estimation: Methods, benchmarks, and challenges," *Sensors*, vol. 25, no. 8, 2025.

[2] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3d human pose estimation: A review," *Comput. Vis. Image Underst.*, vol. 210, Sept. 2021.

[3] R. Neupane, K. Li, and T. Boka, "A survey on deep 3d human pose estimation," *Artificial Intelligence Review*, vol. 58, 11 2024.

[4] C. Chen and D. Ramanan, "3d human pose estimation = 2d pose estimation + matching," *CoRR*, vol. abs/1612.06524, 2016.

[5] A. Aalerud and G. Hovland, "Evaluation of perception latencies in a human-robot collaborative environment," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5018–5023, IEEE, 2020.

[6] G. Scarpellini, P. Morerio, and A. Del Bue, "Lifting monocular events to 3d human poses," in *IEEE CVPRW*, pp. 1358–1368, 2021.

[7] W. Ikeda, M. Hatano, R. Hara, and M. Isogawa, "Event-based egocentric human pose estimation in dynamic environment," in *IEEE ICIP*, pp. 2336–2341, 2025.

[8] G. Goyal, F. D. Pietro, N. Carissimi, A. J. Glover, and C. Bartolozzi, "Moveenet: Online high-frequency human pose estimation with an event camera," *IEEE/CVF CVPRW*, pp. 4024–4033, 2023.

[9] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *IEEE CVPR*, pp. 8877–8886, 2023.

[10] E. Brau and H. Jiang, "3d human pose estimation via deep learning from 2d annotations," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 582–591, 2016.

[11] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *Int. Conference 3DV*, pp. 506–516, 2017.

[12] I. Ramírez, A. Cuesta-Infante, E. Schiavi, and J. J. Pantrigo, "Bayesian capsule networks for 3d human pose estimation from single 2d images," *Neurocomput.*, vol. 379, p. 64–73, Feb. 2020.

[13] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *IEEE/CVF CVPR*, p. 3420–3430, IEEE, June 2019.

[14] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Diffpose: Toward more reliable 3d pose estimation," in *IEEE/CVF CVPR*, pp. 13041–13051, 2023.

[15] M. Moro, G. Marchesi, F. Hesse, F. Odone, and M. Casadio, "Markerless vs. marker-based gait analysis: A proof of concept study," *Sensors*, vol. 22, no. 5, 2022.

[16] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng, "Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild," *Int. J. Comput. Vision*, vol. 129, p. 703–718, Mar. 2021.

[17] S. A. Ghasemzadeh, A. Alahi, and C. De Vleeschouwer, "Mpl: Lifting 3d human pose from multi-view 2d poses," in *ECCV 2024 Workshops*, (Berlin, Heidelberg), p. 36–52, Springer-Verlag, 2025.

[18] P. T. D. Hardy, S. Dasmahapatra, and H. Kim, "Optimising 2d pose representations: Improving accuracy, stability and generalisability within unsupervised 2d-3d human pose estimation," in *Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production*, CVMP '23, (New York, NY, USA), Association for Computing Machinery, 2023.

[19] S. Mehraban, Y. Qin, and B. Taati, "Evaluating recent 2d human pose estimators for 2d-3d pose lifting," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5, 2024.

[20] T.-H. Hoang, M. Zehni, H. Phan, D. M. Vo, and M. N. Do, "Improving the robustness of 3d human pose estimation: A benchmark dataset and learning from noisy input," in *IEEE/CVF CVPRW*, pp. 113–123, 2024.

[21] S. Lee, Y. Hwang, and J. T. Lee, "Learning 2d human poses for better 3d lifting via multi-model 3d-guidance," in *ACCV*, pp. 3344–3361, December 2024.

[22] S. Mehraban, V. Adeli, and B. Taati, "Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network," in *IEEE/CVF WACV*, pp. 6905–6915, 2024.

[23] G. Goyal, P. C. Thuong, A. Glover, M. Mizuno, and C. Bartolozzi, "Graphenet: Event-driven human pose estimation with a graph neural network," in *IEEE/CVF ICCV*, pp. 4665–4674, 2025.

[24] E. Calabrese, G. Taverni, C. Awai Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, and T. Delbruck, "Dhp19: Dynamic vision sensor 3d human pose dataset," in *CVPRW*, June 2019.

[25] S. Zou, C. Guo, X. Zuo, S. Wang, P. Wang, X. Hu, S. Chen, M. Gong, and L. Cheng, "EventHPE: Event-based 3d human pose and shape estimation," in *IEEE/CVF ICCV*, pp. 10976–10985, 2021.

[26] J. He, Z. Zeng, X. Li, and C. Fan, "Evtranspose: Towards robust human pose estimation via event camera.," *Electronics (2079-9292)*, vol. 14, no. 6, 2025.

[27] S. Tang, H. Lv, X. Li, Y. Zhao, Y. Zhang, and Y. Feng, "Event camera-based human pose estimation via hybrid spiking-point cloud neural architecture," *Results in Engineering*, p. 106771, 2025.

[28] V. Rudnev, V. Golyanik, J. Wang, H.-P. Seidel, F. Mueller, M. Elgharib, and C. Theobalt, "Eventhands: Real-time neural 3d hand pose estimation from an event stream," in *IEEE/CVF ICCV*, pp. 12385–12395, 2021.

[29] Y. Kim, H. Cho, and K.-J. Yoon, "From sharp to blur: Unsupervised domain adaptation for 2d human pose estimation under extreme motion blur using event cameras," in *IEEE/CVF ICCV*, pp. 9406–9417, 2025.

[30] L. Gava, M. Monforte, C. Bartolozzi, and A. Glover, "How late is too late? a preliminary event-based latency evaluation," in *Int Conf. on EBCCSP*, pp. 1–4, 2022.

[31] G. Goyal, F. di Pietro, A. Glover, and C. Bartolozzi, "Event-human3.6m dataset." https://doi.org/10.5281/zenodo.7842598, Apr. 2023.

[32] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Trans. on PAMI*, vol. 36, pp. 1325–1339, jul 2014.

[33] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic dvs events," in *IEEE/CVF CVPRW*, pp. 1312–1321, 2021.

[34] L. Gava, *Real-time Event-based Tracking for Low-Latency Robot Interaction with Complex Dynamic Environments*. PhD thesis, University of Manchester, 2024.

[35] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3d human pose estimation using spatio-temporal networks with explicit occlusion training," 2020.

[36] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. on PAMI*, vol. 43, no. 10, pp. 3349–3364, 2021.

[37] X. Zhang, Q. Bao, Q. Cui, W. Yang, and Q. Liao, "Pose magic: efficient and temporally consistent human pose estimation with a hybrid mamba-gcn network," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI Press, 2025.