International Journal of Computer Vision (2024) 132:2943–2961
https://doi.org/10.1007/s11263-024-02011-z

Event-Based Non-rigid Reconstruction of Low-Rank Parametrized Deformations from Contours

Yuxuan Xue[1,2,3,4] · Haolong Li[1] · Stefan Leutenegger[2] · Jörg Stückler[1]

Received: 27 April 2023 / Accepted: 16 January 2024 / Published online: 26 February 2024

Abstract
Visual reconstruction of fast non-rigid object deformations over time is a challenge for conventional frame-based cameras. In recent years, event cameras have gained significant attention due to their bio-inspired properties, such as high temporal resolution and high dynamic range. In this paper, we propose a novel approach for reconstructing such deformations using event measurements. Under the assumption of a static background, where all events are generated by the motion, our approach estimates the deformation of objects from events generated at the object contour in a probabilistic optimization framework. It associates events to mesh faces on the contour and maximizes the alignment of the line of sight through the event pixel with the associated face. In experiments on synthetic and real data of human body motion, we demonstrate the advantages of our method over state-of-the-art optimization and learning-based approaches for reconstructing the motion of human arms and hands. In addition, we propose an efficient event stream simulator to synthesize realistic event data for human motion.

Keywords Event cameras · Non-rigid reconstruction · Human motion reconstruction

1 Introduction

The capturing and 3D reconstruction of real-world scenes is an important field in computer vision with applications in VR/AR or robotics. Tracking and reconstructing non-rigid objects poses challenges due to the high dimensionality of the inference problem when reconstructing complex deforming shapes. Existing methods for non-rigid reconstruction from RGB(-D) (Salzmann & Fua, 2009; Ngo et al., 2016; Yu et al., 2015; Bregler et al., 2000; Dai et al., 2014; Garg et al., 2013; Sidhu et al., 2020; Lamarca et al., 2021) have limitations, particularly in scenarios with fast motion or low lighting conditions due to limits in temporal resolution or motion blur. Event-based cameras, which offer advantages like high dynamic range and low latency, have the potential to excel in non-rigid tracking tasks in dark scenes. Despite their benefits, only a few studies have addressed the problem of non-rigid reconstruction with event cameras.

Event cameras (Lichtsteiner et al., 2008) offer a considerable number of advantages in computer vision tasks over conventional cameras, such as low latency, high dynamic range and virtually no motion blur (Gallego et al., 2022). Unlike conventional frame-based cameras that capture images at a fixed rate, event cameras asynchronously measure per-pixel brightness change, and output a stream of events that encode the spatio-temporal coordinates of the brightness change and its polarity. This measurement principle avoids the limitations of motion blur in frame-based cameras and enables high temporal resolution. In recent years, a significant amount of research has focused on developing event-based approaches for various computer vision applications (Gallego et al.,

Communicated by Guang Yang.

✉ Yuxuan Xue
yuxuan.xue@uni-tuebingen.de

Haolong Li
haolong.li@tuebingen.mpg.de

Stefan Leutenegger
stefan.leutenegger@tum.de

Jörg Stückler
joerg.stueckler@tuebingen.mpg.de

[1] Max Planck Institute for Intelligent Systems, Embodied Vision Group, Max-Planck-Ring 4, 72076 Tübingen, Baden-Württemberg, Germany

[2] Technical University of Munich, Boltzmannstr. 3, 85748 Munich, Bavaria, Germany

[3] University of Tübingen, Maria-von-Linden Str. 6, 72076 Tübingen, Baden-Württemberg, Germany

[4] Tübingen AI Center, Maria-von-Linden Str. 6, 72076 Tübingen, Baden-Württemberg, Germany

2022) including optical flow estimation (Bardow et al., 2016; Gehrig et al., 2021; Gallego et al., 2018; Stoffregen & Kleeman, 2019; Zhu et al., 2018), visual-inertial odometry(VIO) (Kim et al., 2016; Rebecq et al., 2017; Vidal et al., 2018; Bryner et al., 2019), video reconstruction (Rebecq et al., 2019, 2021), object pose estimation (Li & Stueckler, 2021).

While several approaches for event cameras have been proposed for aforementioned computer vision tasks, only little work has been devoted to non-rigid reconstruction. Recently, Nehvi et al. (2021) proposed a non-rigid tracking approach using a differentiable generative event model to simulate event measurements which are compared with the actual measurements in an optimization framework. Differently to this method, we explicitly reason on the association of events to elements in the 3D geometry of the object. Rudnev et al. proposed EventHands (Rudnev et al., 2021), a learning-based framework trained on synthetic event data to reconstruct hand motion. As learning is fully supervised, the method requires annotated data for training and can be limited to the data domain seen during training.

In this paper, we present a novel non-rigid reconstruction approach for event cameras. Our algorithm takes event streams and an initial pose guess as input and outputs the reconstructed object pose parameters, assuming a low-dimensional parameterized shape template of a deforming object (i.e. hand and body model). To achieve this, we propose a novel optimization-based method based on expectation maximization (EM). Our method models event measurements at contours of the non-rigid 3D shape model in a probabilistic way to estimate the association likelihood of events to mesh faces and maximize the measurement likelihood. We evaluate our approach on synthetic and real data sequences, and demonstrate the improvements over the state-of-the-art optimization (Nehvi et al., 2021) and learning-based (Rudnev et al., 2021) methods for hand reconstruction. In summary, our contribution are:

- We propose a novel non-rigid reconstruction approach for event cameras based on expectation maximization. In our experiments, it demonstrates better accuracy than state-of-the-art event-based non-rigid reconstruction approaches while being robust to different level of noise.
- We also develop an efficient event stream simulator for human motion sequences. It supports 5 different data modalities as well as synthesis of various noise sources.

2 Related Work

2.1 Event Camera

As a bio-inspired sensor, event-based cameras (Lichtsteiner et al., 2008; Gallego et al., 2022) capture the logarithmic pixel-wise brightness change asynchronously, mimicking how the retina works in our brain. Thus, the outputs of traditional cameras and event-based cameras are different: traditional cameras acquire the visual information of a scene as a stream of intensity frames at a constant rate, while event-based cameras have no notion of images since each pixel operates independently. Event cameras provide significant advantages over conventional frame-based cameras:

- High temporal resolution: events are measured with microsecond resolution (1 MHz) (Gallego et al., 2022)

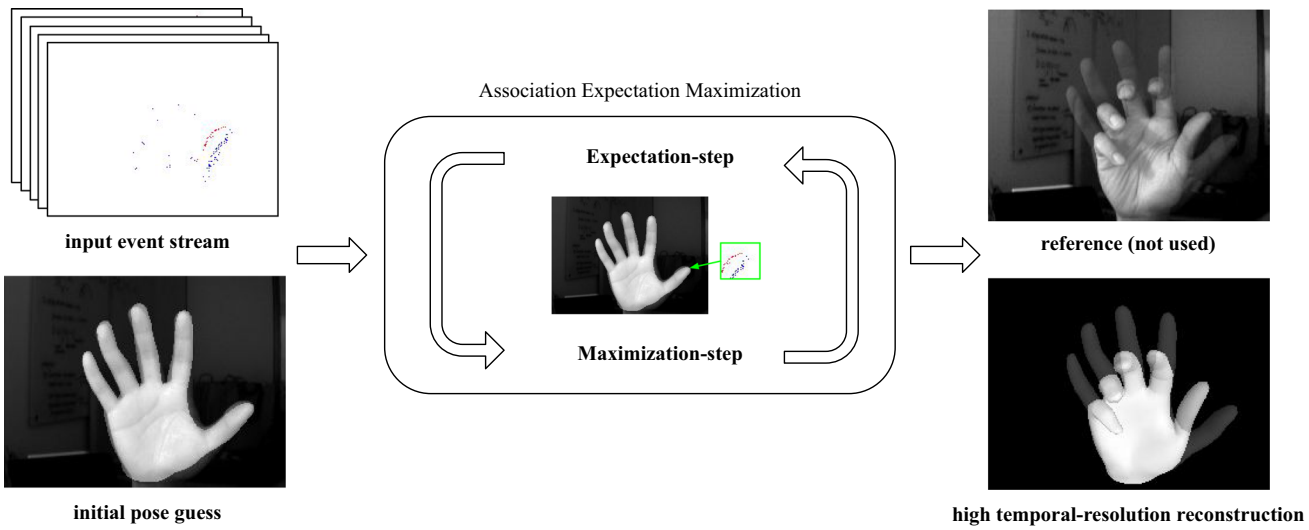


Fig. 1 Our approach performs non-rigid reconstruction with high temporal resolution from measurements of event-based cameras within an expectation-maximization (EM) framework and using an initial pose guess, which can, for instance, be obtained from frame-based cameras

when the brightness at pixel-level changes by a threshold. As comparison, conventional frame-based cameras usually have a frequency of 30 to 60 Hz. Due to the measurement principle, event-based cameras do not suffer from motion blur like frame-based cameras.
- High dynamic range: event cameras have high dynamic range (> 120 dB), which exceeds the 60 dB of typical high-quality frame-based cameras. Therefore, event cameras can adapt to very dark as well as very bright stimuli.
- Low power and storage consumption: event cameras only capture pixel-level brightness change asynchronously, which means it avoids redundant data. Power and storage are only used to process pixel brightness changes.

Since event cameras do not have images as output, the existing computer vision algorithms for conventional cameras are not directly feasible to event-based cameras. Recently, several computer vision tasks have been addressed with event cameras (Gallego et al., 2022). Considering their advantageous features such as low latency, low power consumption, and high dynamic range, event-based cameras have been applied to robotics problems such as optical flow (Bardow et al., 2016; Gehrig et al., 2021; Gallego et al., 2018; Stoffregen & Kleeman, 2019; Zhu et al., 2018), SLAM and VIO (Kim et al., 2016; Rebecq et al., 2017; Vidal et al., 2018; Bryner et al., 2019). Event-based non-rigid reconstruction as pursued in this work has received limited attention from the research community so far (see Sect. 2.3).

2.2 Non-rigid Reconstruction

Reconstruction and tracking of non-rigid shapes is a challenging problem in computer vision. For monocular frame-based cameras, several approaches have been proposed. They can be classified into methods that align shape templates [e.g. (Salzmann & Fua, 2009; Ngo et al., 2016; Yu et al., 2015)] or approaches that use regularizing assumptions such as low-rank approximations to achieve non-rigid structure from motion [e.g. (Bregler et al., 2000; Dai et al., 2014; Garg et al., 2013; Sidhu et al., 2020; Lamarca et al., 2021)]. Using RGB-D cameras simplifies the task due to the availability of dense depth for which several methods have been proposed recently [e.g. (Newcombe et al., 2015; Bozic et al., 2020)].

Human motion reconstruction is a specific type of problem setting which facilitates template-based non-rigid reconstruction. While other deformable objects can exhibit significant shape variations, human bodies possess inherent similarities, allowing for parametrization in low-dimensional shape models (Anguelov et al., 2005; Loper et al., 2015; Pavlakos et al., 2019; Romero et al., 2017). The process of reconstructing a dynamic human body, based on the provided parametric template, is comparable to solving for the pose parameters within the low-rank space. Various studies (Sun et al., 2021; Lin et al., 2021; Aboukhadra et al., 2023; Li et al., 2023, 2021) have focused on estimating the 3D body pose using frame-based cameras.

The previously mentioned methods use frame-based cameras for reconstructing non-rigid deformations. However, these approaches are constrained by the limitations inherent to conventional cameras, such as motion blur and relatively low dynamic range. Our novel event-based non-rigid reconstruction method leverages advantages offered by event cameras.

2.3 Event-Based Non-rigid Reconstruction

Non-rigid reconstruction and tracking with event cameras has only recently attained attention in the computer vision community. Nehvi et al. (2021) propose a differentiable event stream simulator by subtracting renderings of parametrized hand models (Romero et al., 2017). The paper demonstrates the use of the simulator for non-rigid motion tracking from event streams by optimization. However, the tracking performance of Nehvi's method is constrained by the quality of the generated events. Non-robustness in tracking arises when the capturing scenario deviates from a pure black background, resulting in differences between the captured and generated events. Rudnev et al. (2021) propose EventHands, which trains a deep neural network on synthetic event streams to estimate the deformation of a MANO (Romero et al., 2017) hand model. To input the event data into the neural network, they propose to represent the data in local time windows. However, EventHands is limited to synthetic hand training data since there is a lack of labeled real-world event data. Consequently, there can be a performance drop when a distribution shift occurs between test and synthetic training data.

Different to these methods, we propose geometric contour alignment in a probabilistic optimization framework. Our approach, unlike Nehvi's (Nehvi et al., 2021), lifts 2D events to 3D by a contour measurement model and optimizes the object pose parameters using the expectation maximization (EM) algorithm. The EM algorithm associates and aligns events to contours on the 3D shape which results in improved accuracy and robustness. Moreover, compared to EventHands (Rudnev et al., 2021), our approach is optimization-based and does not depend on synthetic training data for deep learning, thus avoiding issues of distribution shift.

3 Background

3.1 Non-rigid Parametric Models

3.1.1 Parametric Body Model

SMPL (Loper et al., 2015) is a widely used parametric model for representing human body shapes and poses. SMPL model begins with a canonical template mesh $\overline{\mathbf{T}}$, and subsequently incorporates shape blend shapes B_S (controlled by shape parameters $\boldsymbol{\beta}$) and pose blend shapes B_P (controlled by pose parameters $\boldsymbol{\theta}$) as additional vertex offsets

$$\begin{aligned} T_C(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \overline{\mathbf{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}), \\ B_S(\boldsymbol{\beta}; \mathcal{S}) &= \sum_{n=1}^{|\boldsymbol{\beta}|} \beta_n \mathbf{S}_n, \\ B_P(\boldsymbol{\theta}; \mathcal{P}) &= \sum_{n=1}^{9K} R_n(\boldsymbol{\theta}) \mathbf{P}_n, \end{aligned} \tag{1}$$

where $\mathbf{S}_n$ and $\mathbf{P}_n$ are shape and pose blend shape deformation matrices. $R_n(\theta)$ is the element in rotation matrices of the pose θ. The resulting mesh T_C is in canonical space, and is transformed into deformation space by Linear Blend Skinning and joints rotation

$$T_P = \left(\sum_{i=1}^{K} w_i \mathbf{G}_i(\boldsymbol{\theta}) \right) T_C, \tag{2}$$

where $\mathbf{G}_i(\boldsymbol{\theta})$ is the transformation matrix of joint i and w_i is the skinning weights. Please refer to Sect. 3 and Fig. 3 in SMPL (Loper et al., 2015) for more details about the parametric blend shapes as well as the skinning. In our work, we can consider SMPL (Loper et al., 2015) as a linear model, which takes pose parameter as input and outputs the given posed mesh.

3.1.2 Parametric Hand Model

MANO (Romero et al., 2017) is a generative hand model that can map the hand pose parameter and shape parameter into a 3D hand mesh. Each hand posture is parameterized by a set of principle components coefficients that map a differentiable low-dimensional manifold. In our scenario, we use the full 45-dimensional PCA parameters of MANO to formulate the pose space.

3.1.3 Parametric Expressive Model

SMPL-X (Pavlakos et al., 2019) is an expressive parametric human model, which models shape and pose of the human body using SMPL (Loper et al., 2015), hand pose using MANO (Romero et al., 2017), and facial expression using FLAME (Li et al., 2017; Feng et al., 2021). The body pose is represented by 3-DoF orientations of 21 joints, hand pose is controlled by 45 PCA parameters, and facial expression is controlled by 10 PCA parameters in expression space. The SMPL-X model can map the body pose, hand pose, and the facial parameters into an expressive and detailed 3D body mesh.

3.2 Expectation Maximization

The Expectation Maximization (EM) algorithm is an iterative optimization method to estimate the maximum likelihood or maximum a posteriori parameters of a probabilistic model. The EM algorithm is a powerful approach for handling missing or unobserved data. In the maximum likelihood case, the problem with missing data can be formulated as

$$\ln p(\mathbf{x} \mid \boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) \right\}, \tag{3}$$

where $\mathbf{x}$, $\boldsymbol{\theta}$, $\mathbf{z}$ represent observation, model parameters, and unobserved latent variables, respectively.

The knowledge of latent variables $\mathbf{z}$ is only given by the posterior distribution $p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta})$. Because the complete-data log likelihood is not available due to the missing data (unobserved latent variables $\mathbf{z}$), it is determined by an expected value under the posterior distribution of the latent variable. The expectation of the complete-data log marginal likelihood is written as

$$\mathcal{Q}\left(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}\right) = \sum_{\mathbf{z}} p\left(\mathbf{z} \mid \mathbf{x}, \overline{\boldsymbol{\theta}}\right) \ln p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}), \tag{4}$$

which is maximized in the M-step to revise the new model parameter estimate $\boldsymbol{\theta}^{new}$. Note that $\overline{\boldsymbol{\theta}}$ is the current estimate of model parameters from the previous iteration, which is treated as constant in optimization. In the E-step, the optimal posterior distribution $p\left(\mathbf{z} \mid \mathbf{x}, \overline{\boldsymbol{\theta}}\right)$ is determined based on the latest parameter estimate. Both E- and M-steps are alternated for the optimization. In our approach, we model the events association w.r.t. the given template as fully unobserved data, i.e., the problem is a Missing Completely At Random (MCAR) problem. We detail our EM-formulation for event-based reconstruction in Sect. 5.1.

4 Non-rigid Event Stream Simulator

For data generation and evaluation, we develop an efficient event stream simulator capable of generating synthetic event data for temporally deforming objects. Our simulator goes beyond solely producing events and can also generate RGB

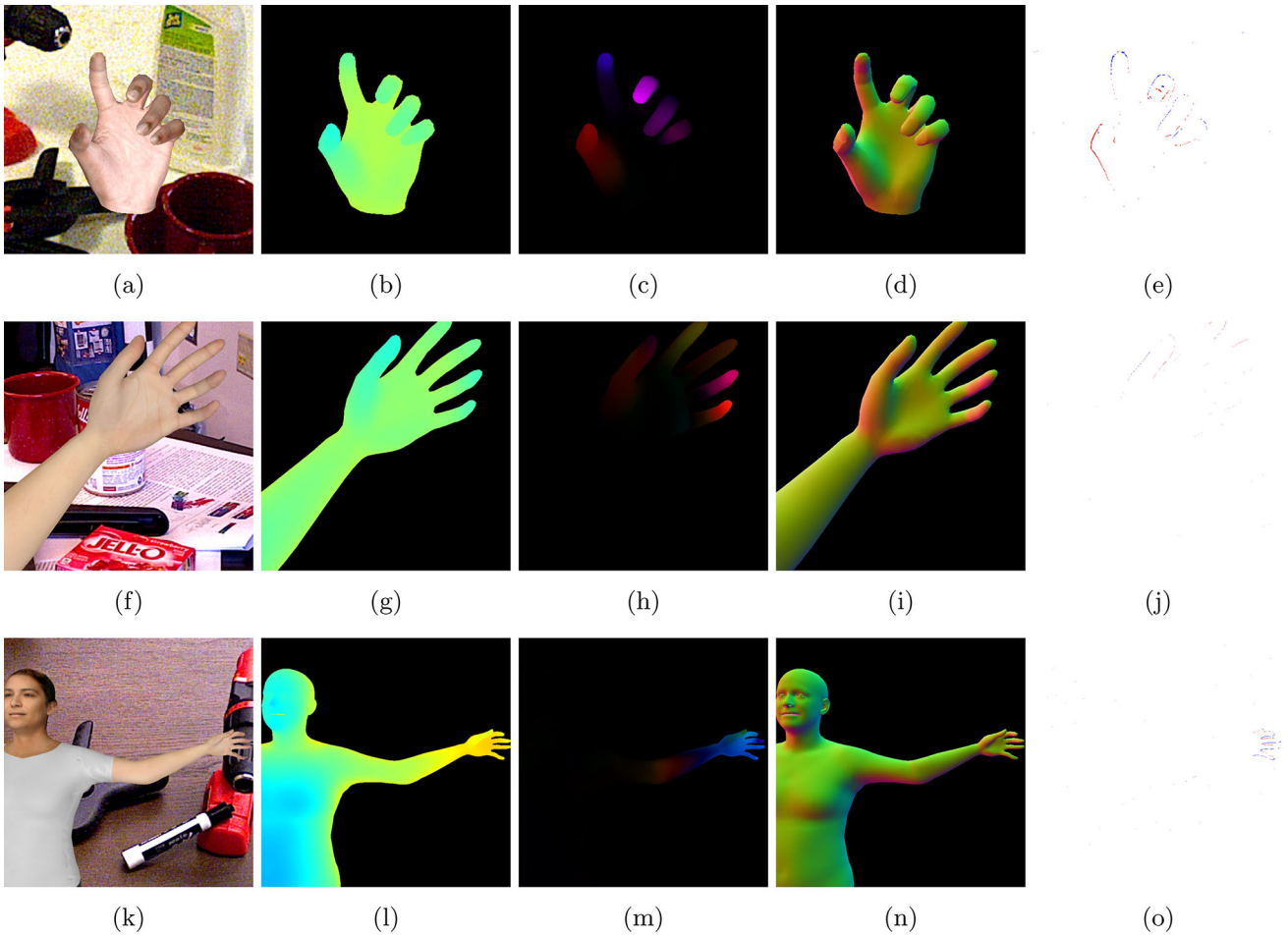


Fig. 2 Data modalities for MANO (Romero et al., 2017) hand model (**a**–**e**), SMPL-X (Pavlakos et al., 2019) hand model (**f**–**j**), and SMPL-X (Pavlakos et al., 2019) arm and hand model (**k**–**o**). Data include (**a**, **f**, **k**) RGB image, (**b**, **g**, **l**) depth map, (**c**, **h**, **m**) motion field, (**d**, **i**, **n**) normal map, (**e**, **j**, **o**) accumulated events in 1/30 s. Background images in RGB images are randomly selected from YCB video dataset (Xiang et al., 2018)

images, depth images, optical flow, and surface normals based on a sequence of pose parameters.

4.1 Event Generation Model

Unlike RGB cameras which capture absolute brightness for each pixel at a fixed frame rate, event cameras record logarithmic pixel-level brightness change asynchronously. To simulate the event at time t_i, we calculate the absolute logarithmic brightness at each pixel $\mathbf{u_i}$, denoted as $\mathcal{L}(\mathbf{u_i}, t_i)$, and compare it with the logarithmic brightness value of the last sampled image at time t_{i-1}. The polarity p of the event is

$$p(\mathbf{u}_i, t_i) = \begin{cases} +1 & \text{if } \mathcal{L}(\mathbf{u}_i, t_i) - \mathcal{L}(\mathbf{u}_i, t_{i-1}) \geq C^+, \\ -1 & \text{if } \mathcal{L}(\mathbf{u}_i, t_{i-1}) - \mathcal{L}(\mathbf{u}_i, t_i) \geq C^-, \end{cases} \tag{5}$$

where C^+ and C^- are positive and negative contrast threshold, respectively. If the logarithmic brightness change is less than the corresponding contrast threshold, no event is generated at pixel $\mathbf{u}_i$.

4.2 Simulation Approach

Our simulator takes a sequence of pose parameters of body and hand, the facial expression parameters, and the simulation time as inputs and simulates event stream, RGB image, depth map, optical flow, and normal map (see Fig. 2). Similar as other event stream simulators (Rebecq et al., 2018; Nehvi et al., 2021; Rudnev et al., 2021), our simulator assumes that the pose and expression parameters change linearly between two consecutive inputs of the sequence.

An example of a simulated data stream for the MANO (Romero et al., 2017) model is shown in Fig. 2 in (a–e). Note that the MANO hand model only contains the hand but no arm. We visualize the simulated data stream for the SMPL-X (Pavlakos et al., 2019) hand model in an example in Fig. 2

in (f–j). It only contains the motion of the hand. However, attaching the arm to the hand makes it more realistic. The simulated data modalities of SMPL-X (Pavlakos et al., 2019) arm and hand motion are visualized in Fig. 2 in (k–o).

4.3 Adaptive Sampling

An essential advantage of event-based cameras is the independent representation of the visual signal at every pixel. Unlike the intensity images which have a fixed frame rate, the events are measured in an asynchronous fashion on the pixel array. Inspired by ESIM (Rebecq et al., 2018), we adapt the sampling rate based on the predicted change of the visual signal using the optical flow at each pixel. The color image is rendered in the calculated sampling rate. The events since the last color image are simulated using the event generation principle in Sect. 4.1.

To simulate the optical flow, we project the 3D movement of each mesh face onto the 2D image plane. The idea of adaptive sampling based on optical flow is to ensure that the maximum displacement of any pixel on the image plane between consecutive rendered frames is bounded:

$$t_{k+1} = t_k + \lambda_v |\mathcal{V}|_m^{-1}, \tag{6}$$

where $|\mathcal{V}|_m = \max_{x\in\Omega} |\mathcal{V}(x; t_k)|$ is the maximum magnitude of the motion vector across the image plane at time t_k. We use $\lambda_v \leq 1$ to manually control the render rate and adjust it using real event data. For more details regarding the adaptive sampling, we kindly ask readers to refer to ESIM (Rebecq et al., 2018).

4.4 Noise Synthesis

Measurements of event cameras are often noisy. To make the simulated data more realistic, we also inject noise in the simulated event stream. As in ESIM (Rebecq et al., 2018), we sample the contrast threshold from a normal distribution with standard deviation σ for each pixel at every sampling step to add uncertainty to the event generation. To simulate salt-and-pepper noise on the background, we sample the probability of each pixel to generate an outlier event from a uniform distribution in [0, 1] and compare it with a predefined threshold. If the probability exceeds the threshold, a noise event is generated. We then sample the timestamp of the noise events uniformly in $[t_{i-1}, t_i]$. For more details on how to adjust the threshold for obtaining a similar amount of salt-and-pepper noise as real event cameras, please refer to EventHands (Rudnev et al., 2021).

4.5 Comparison

We present a comparative analysis of our proposed simulator with existing event stream simulators (Nehvi et al., 2021; Rebecq et al., 2018; Rudnev et al., 2021) in Table 1. Our simulator offers significant advancements over Nehvi's simulator (Nehvi et al., 2021), as it is capable of generating optical flow for deforming objects and achieves acceleration through parallel processing of all pixels within a frame. To assess the efficiency of our simulator, we conducted experiments by simulating the same hand motion using the MANO model (Romero et al., 2017) on an NVidia RTX-2080Ti GPU. The results indicate that our simulator (47.17 s) outperforms Nehvi's (3717.56 s) by a remarkable factor of 78 in terms of run-time. In comparison to ESIM (Rebecq et al., 2018), our simulator extends its capabilities to simulate events for non-rigid human body motion. Furthermore, our simulator utilizes the adaptive sampling strategy to avoid redundant calculations for small motion. In contrast, EventHands (Rudnev et al., 2021) samples image frames at a fixed rate of 0.001 s, regardless of the actual motion characteristics. This adaptive sampling approach adopted in our simulator improves realism.

5 Event-Based Non-rigid Reconstruction from Contours

Our event-based reconstruction method estimates deformations of parameterized non-rigid objects assuming a static background. Typically, for deforming texture-less objects such as hands or human bodies, the majority of events is generated at the contour between the object and the background (e, j, o in Fig. 2). Hence, we formulate the reconstruction problem in a probabilistic way using a contour measurement model for the events as in Fig. 3. Assuming a known initial state, we optimize for the pose parameters of the parametric object model incrementally from the event stream.

5.1 Expectation Maximization Framework

We formulate the 4D reconstruction problem as maximum-a-posteriori (MAP) estimation of the model parameters $\boldsymbol{\theta}$ given the event observations $\mathbf{x}$ from the event camera

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{x} \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}), \tag{7}$$

where $p(\boldsymbol{\theta})$ is a prior on the parameters obtained with a constant-velocity motion model from the parameters of the previous event buffer. In practice, we aggregate a fixed number of events into an event buffer, assume that the event observations at each pixel are independent from each other,

Table 1 Comparison between our event simulator and other event simulators in different properties

	Objects	Modalities	Adaptive sampling	Parallelism
Our simulator	Body, hand, face	5	✓	✓
ESIM (Rebecq et al., 2018)	Rigid	5	✓	✗
Nehvi et al. (2021)	Hand	1	✗	✗
EventHands (Rudnev et al., 2021)	Hand	1	✗	✓

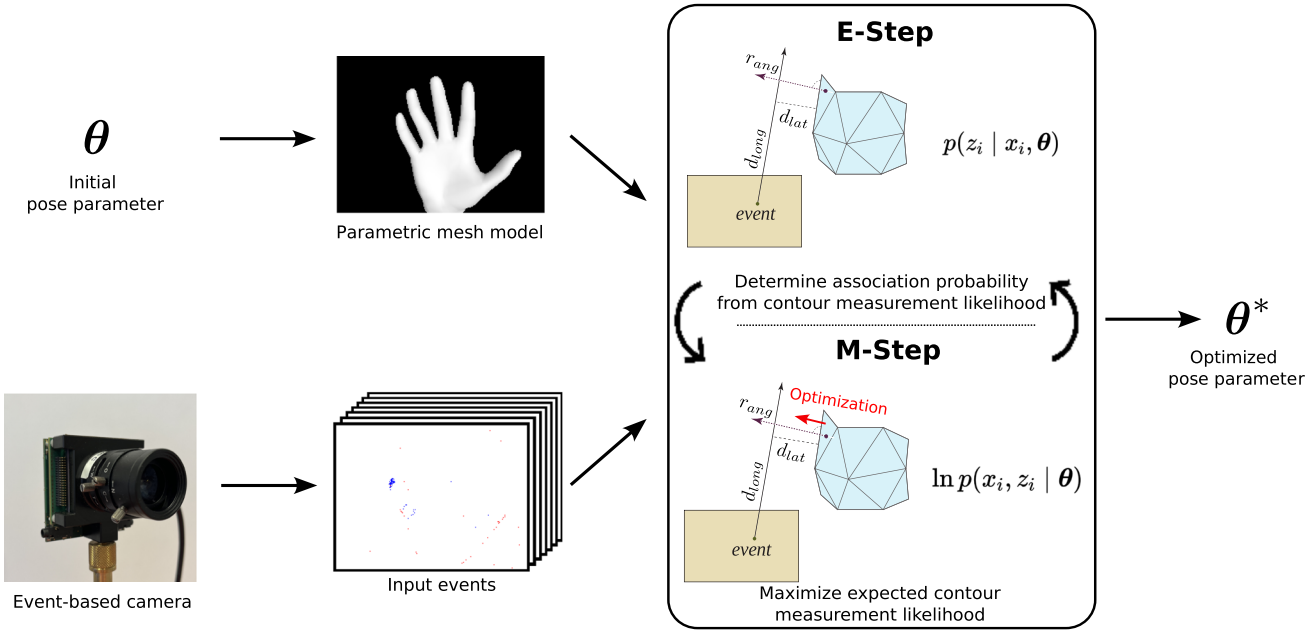


Fig. 3 Our approach reconstructs non-rigid deformation states of objects from event streams of event cameras within an expectation maximization (EM) framework given the initialized template. In the E-step, the association probability of events to contour mesh faces is estimated from the contour measurement likelihood. In the M-step, the expected value of the measurement likelihood over the the association probability is maximized for pose parameter $\boldsymbol{\theta}^*$

and optimize over this event buffer

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln p(x_i \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}), \tag{8}$$

where N is the number of aggregated events in an event buffer $\mathbf{x}$ and x_i denotes the 2D location of the event i on the image plane. By this aggregation, computation can be parallelized in the event buffer more efficiently with a trade-off against the temporal resolution of the data. The analytical solution of the MAP is difficult to obtain because there is no observable relation between measurement $\mathbf{x}$ and model parameters $\boldsymbol{\theta}$ available without further knowing how the events are generated from the model. To solve the MAP in Eq. 8, we introduce a set of latent variables $\mathbf{z}$ as in Eq. 3.

Since our background is static and the objects are textureless, most of the event measurements are generated at the contour of the deforming object. We thus assume that the events in x_i are generated at a point on the observed contour of the object. However, the point on the mesh which generated an event is unknown and needs to be estimated as well. We therefore introduce the latent variable $z_i = j$ which represents the association between the event at x_i and a mesh face of the object with index j on which the event is generated,

$$\ln p(x_i \mid \boldsymbol{\theta}) = \ln \sum_{i=1}^{F} p(x_i, z_i = j \mid \boldsymbol{\theta}). \tag{9}$$

We use the expectation-maximization (EM) framework in Eq. 3 to find the model parameters with the latent association,

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln \left(\sum_{j=1}^{F} p(x_i \mid z_i = j, \boldsymbol{\theta}) \right) + \ln p(\boldsymbol{\theta}), \tag{10}$$

where F is the number of mesh faces and for which we assume a uniform distribution for $p(z_i \mid \boldsymbol{\theta})$. The optimiza-

tion of the next event buffer is initialized with the parameters from the previous buffer. Here we assume that the events in an event buffer are conditionally independent of each others given the pose parameters. We provide a detailed derivation of Eq. 10 from Eq. 7 in the supplementary material.

The expectation of the log marginal likelihood in Eq. 4 is formulated as

$$\sum_{i=1}^{N}\sum_{j=1}^{F} q(z_i = j) \cdot \ln p\left(x_i \mid z_i = j, \boldsymbol{\theta}\right) + \ln p(\boldsymbol{\theta}), \tag{11}$$

with the probabilistic belief on the latent association variables $q(z_i = j)$ using a variational approximation given the current estimate of parameters $\overline{\boldsymbol{\theta}}$ from the previous iteration,

$$q(z_i = j) = p(z_i = j \mid x_i, \overline{\boldsymbol{\theta}}). \tag{12}$$

In the E-step of our EM-framework, we update the probabilistic belief on the latent association variables $q(z_i = j)$. In the M-step, the parameters are updated by maximizing the expected log posterior with the probabilistic, i.e. soft, data association from the E-step,

$$\boldsymbol{\theta} \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N}\sum_{j=1}^{F} q(z_i = j) \ln p(x_i \mid z_i = j, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}). \tag{13}$$

The posterior includes terms for the expected value of the measurement likelihood under the association likelihood $q(z_i)$ and a prior term on the parameters. In the following, we explain the concrete form of the EM steps in detail.

5.2 Data Association

Ideally we can use mesh rasterization to find the association of pixels to all mesh faces that intersect the line of sight through each pixel. Due to limits in the image resolution, initial inaccuracies of the shape parameters during optimization, and complex mesh topologies which allow for multiple layers being intersected by the line of sight, the rasterization often misses the correct association of contour mesh faces with event pixels. For instance, if the shape estimate is off, the contour mesh face could be rendered a few pixels off the observed event location. If fingers are bent in front of the palm, events observed on the contour of the finger might hit palm mesh faces, but miss mesh faces on the finger which generated the events.

The EM-framework requires to quantify the probability of associating a mesh face with an event. Intuitively, the closer the mesh face to the event's unprojection ray, the higher the probability is that it causes the event. Inspired by SoftRasterizer (Liu et al., 2019), we formulate the contour measurement likelihood as

$$p(x_i \mid z_i = j, \boldsymbol{\theta}) \propto \sigma\left(\delta_j^i \frac{d_{\text{lat}}^2(i,j)}{\alpha}\right) \exp\left(-\frac{d_{\text{long}}(i,j)}{\beta}\right) \exp\left(-\frac{r_{\text{ang}}(i,j)}{\gamma}\right), \tag{14}$$

with lateral distance d_{lat}, longitudinal distance d_{long} and angular error r_{ang} between the line of sight through event x_i and the mesh face f_j, and sigmoid function σ. The angular error r_{ang} measures the deviation of the direction of the line of sight from being orthogonal to the normal of the mesh face. Hyperparameters α, β, and γ are used to control the sharpness of the individual terms for the probability distribution.

The lateral distance d_{lat} is the distance between the line of sight and the closest edge of the mesh face. The sign indicator is defined as $\delta_j^i := \{+1, \text{if } x_i \in f_j; -1, \text{otherwise}\}$. We use a maximal lateral distance threshold $\tau_{lateral}$ to reject outlier events due to noise and unmodelled effects. As the longitudinal distance d_{long}, we determine the projected distance between the event pixel and the mesh face center on the line of sight. As sketched above, the line of sight may intersect multiple mesh faces on the deformed object. The longitudinal distance gives higher likelihood to the mesh face closer to the camera. The line of sight through an event caused by the object contour should be approximately orthogonal to the normal of the corresponding mesh face. For this, we assume that our object mesh is a closed watertight mesh with sufficient resolution. The angular error r_{ang} is thus computed by the absolute dot product between the unit direction vector of the line of sight and the face normal.

5.3 E- and M- Step

In the E-step, we determine the latent association likelihood using the measurement likelihood $p(x_i \mid z_i = j, \boldsymbol{\theta})$ based on Bayes' theorem (Davies, 1988)

$$\begin{aligned} q(z_i = j) &= p(z_i = j \mid x_i, \boldsymbol{\theta}) \\ &= \frac{p(x_i \mid z_i = j, \boldsymbol{\theta}) p(z_i = j \mid \boldsymbol{\theta})}{\sum_{j'} p(x_i \mid z_i = j', \boldsymbol{\theta}) p(z_i = j' \mid \boldsymbol{\theta})} \\ &= \frac{p(x_i \mid z_i = j, \boldsymbol{\theta})}{\sum_{j'} p(x_i \mid z_i = j', \boldsymbol{\theta})}, \end{aligned} \tag{15}$$

where we assume that the association variable $z_i = j$ does not dependent stochastically on the mesh model θ without the measurement variable x_i. I.e., we assume that the probability $p(z_i = j \mid \boldsymbol{\theta})$ has a uniform distribution.

For the M-step, we evaluate the measurement likelihood as

$$p(x_i, z_i = j \mid \boldsymbol{\theta}) = p(x_i \mid z_i = j, \boldsymbol{\theta}) p(z_i = j \mid \boldsymbol{\theta}) \propto \sigma \left(\delta_j^i \frac{d_{\text{lat}}^2(i,j)}{\alpha} \right) \exp\left(-\frac{r_{\text{ang}}(i,j)}{\gamma} \right), \quad (16)$$

where $p(z_i = j \mid \boldsymbol{\theta})$ is neglected due to the uniform distribution assumption. Here we do not include the longitudinal distance term as this would pull the object towards the camera. Ideally, this term should assign a constant probability to mesh faces on the same occlusion layer. Notably the ideal term does not depend continuously on the shape parameters, but is difficult to calculate. Instead, if included in the M-step, our approximative Gaussian term for the E-step would falsely incentivize shape parameters for which the mesh intersects the line of sight closer to camera. Note that in the E-step, the longitudinal distance is needed to disfavor associations of events to occluded mesh faces. The angular error term in the M-step encourages the alignment of events with contours. For scenes with many outlier events (e.g. textured objects), we choose a larger value of γ. The prior term for the M-step is a constant velocity prior on the parameters, i.e., $\ln p(\boldsymbol{\theta}) = k \|\mathbf{v} - \mathbf{v}'\|_2^2$, $\mathbf{v} = \frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{\Delta t}$, where $\boldsymbol{\theta}'$ and $\mathbf{v}'$ are the parameters and velocity for the previous event buffer and Δt is the time difference between the two event buffers. We alternate E-step and M-step until convergence. When a new event buffer is available, we initialize $\boldsymbol{\theta}$ based on the current estimate of $\mathbf{v}$.

6 Experiments

We evaluate and demonstrate our event-based non-rigid reconstruction approach on synthetic and real sequences using MANO and SMPL-X object models, involving random motions and various background textures. We provide qualitative and quantitative results, comparing with state-of-the-art baselines. At the end of the section, an evaluation of the robustness against noisy events and initial poses, an ablation study for the terms in our E- and M-steps, and results for a hard-EM variant are available. Please also refer to the supplemental video for more qualitative results.

6.1 Experiment Setting

6.1.1 Implementation Details

We accumulate noisy events in event buffers and optimize the shape parameters for each event buffer sequentially. Similar to Vidal et al. (2018), we accumulate buffers with a fixed number of events, therefore choosing their temporal length adaptively. For our real captured data sequences, we accumulate 100 events per buffer. For synthetic data generated by our simulator, we stack 300 events into each buffer. We simulate a Prophesee camera with image size 1280×720 pixels. The event contrast threshold is 0.5. We use the pinhole camera model for the event camera and assume the camera intrinsics are calibrated. Our algorithm optimizes for the pose parameters of the MANO hand model (Romero et al., 2017) and the SMPL body model (Loper et al., 2015; Pavlakos et al., 2019). In case of the hand model, the pose parameters are in PCA space and the MANO modeling approach reconstructs the vertex offsets, which are used together with the canonical pose vertices to generate the posed mesh. For the SMPL body model, the pose parameters represent the joint orientations. We apply Linear Blend Skinning (LBS) together with optimized pose parameters to recover the posed mesh.

6.1.2 Datasets

We generate synthetic datasets of sequences with three types of different objects, namely the MANO (Romero et al., 2017) hand (Fig. 6a), the SMPL-X (Loper et al., 2015; Pavlakos et al., 2019) hand (Fig. 7a), and the combined SMPL-X (Loper et al., 2015; Pavlakos et al., 2019) arm & hand (Fig. 9a). MANO hand sequences are generated with a single hand mesh at a fixed position and orientation. We vary the full 45-dimensional pose parameter space to generate varying hand poses. For SMPL-X hand sequences, the hand is attached to the whole human body which prevents observing the inside of the hand mesh. We vary the first 6 principal component pose parameters to simulate time-varying and realistic hand deformations. In the SMPL-X arm & hand sequences, we synthesize the arm motion by the 3-DoF rotation of the elbow joint and the hand motion by the 6 principal pose parameters. We use the proposed event simulator with adaptive sampling rate to generate the synthetic sequences for the different object models.

For each sequence, the background image is randomly chosen from a texture-rich indoor scene in the YCB video dataset (Xiang et al., 2018). To introduce noise into the event generation process, we sample the contrast threshold of each pixel from a Gaussian distribution with standard deviation 0.0004. The threshold of salt-and-pepper noise is 10^{-5}.

In Fig. 4, we provide a histogram over the event buffer time lengths for all SMPL-X hand sequences. It can be observed that due to the varying degree of image motion, the buffers have varying temporal length. The median length of the event buffers is approx. 0.007 s, which corresponds to approx. 142 Hz.

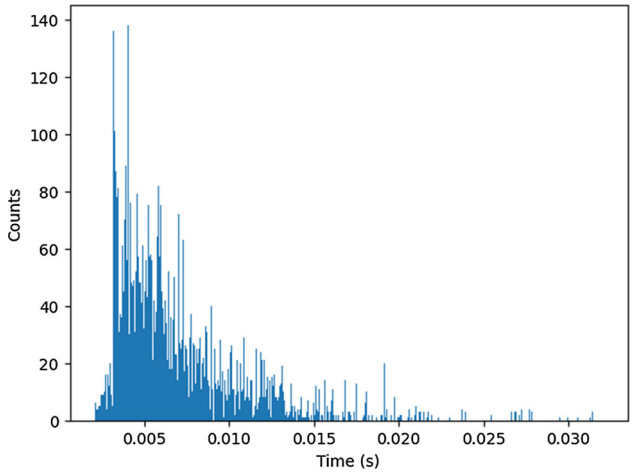


Fig. 4 Histogram of the temporal length of event buffers in the SMPL-X hand sequences

6.1.3 Evaluation Metrics

For the synthetic data, 3D ground-truth positions for all joints and mesh vertices as well as pose parameters are known. We evaluate using the Mean Per Joint Position Error (MPJPE (von Marcard et al., 2018)), the percentage of correct 3D Joints (3D-PCK (Mehta et al., 2017)), and the area under the PCK-curve (AUC (Mehta et al., 2017)) with thresholds ranging from 0 to 50 mm. For hand sequences, we consider the 15 hand skeleton joints. For arm & hand sequences, only the forearm and the hand have motion. Thus, we consider one wrist joint and 15 hand joints.

6.2 Pose Initialization

Our approach assumes the initial pose and shape parameters are known. For real data, we used MeshGraphormer (Lin et al., 2021) to infer MANO mesh model parameters from a grayscale image. We also adopt MeshGraphormer for initializing Nehvi's method (Nehvi et al., 2021). We minimize the Chamfer distance between the predicted hand mesh and the PCA-parametrized MANO hand mesh to optimize shape and pose parameters of the captured hand. Finally, we fix the pose parameters of the hand, and manually fine-tune the global rotation and translation of the mesh model by visual alignment between the rendered 2D hand image and captured hand image.

6.3 Hyperparameter Tuning

We use Optuna (Akiba et al., 2019) to tune hyperparameters of our approach and Nehvi's method (Nehvi et al., 2021). The hyperparameters in our work comprise sharpness control parameters (α, β, γ), early stopping threshold in the optimization, expectation update threshold, and the outlier distance threshold. The hyperparameters in Nehvi's method are the contrast threshold C, the smoothness control weight w, and weights of individual loss terms.

For each scenario, we have 10 random training sequences to tune the hyperparameters. We use the MPJPE as the metric of the loss function. Optuna minimizes the MPJPE error to find the hyperparameters. We use different settings of hyperparameters for the MANO and the SMPL-X model.

6.4 Quantitative Evaluation

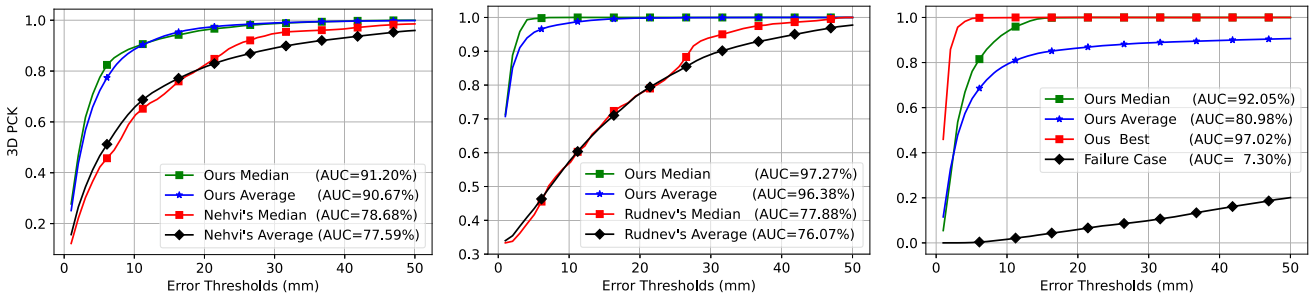
We compare our approach quantitatively with the state-of-the-art event-based non-rigid object reconstruction methods: Nehvi's optimization-based approach (Nehvi et al., 2021) evaluates using the MANO hand model, while Rudnev's approach (Rudnev et al., 2021) is designed for the SMPL-X hand model. To the best of our knowledge, previous event-based reconstruction approaches have not been demonstrated on combined arm & hand motion of a SMPL-X model. Hence, we only provide results for our method on these sequences.

For synthetic MANO hand sequences, we use the MANO (Romero et al., 2017) hand model as the parametric mesh template. In the experiments, we initialize the optimized parameters with the ground-truth pose parameters and evaluate, how well the approach can keep track of the hand deformation. Our approach reconstructs the 45-dimensional pose parameter. We report quantitative results MPJPE and AUC on these sequences in Table 2. Similar to our approach, Nehvi's method is optimization-based and requires the initial parameters of the mesh template. To ensure a fair comparison, we use Optuna (Akiba et al., 2019) to tune hyperparameters in Nehvi's method and our method. We observe that our approach is about 2.5-times more accurate than Nehvi's method. We also show the 3D-PCK curve of both approaches in Fig. 5a. Apparently, our method has higher AUC than Nehvi's method. Results in Table 2 and Fig. 5a demonstrate that our method outperforms Nehvi's method clearly in this dataset.

In the SMPL-X hand sequences, the hand is attached to a human body model. Here, our approach reconstructs the 6-dimensional pose parameters, which is consistent with the evaluation conducted in Rudnev's method (Rudnev et al., 2021). We report quantitative results in Table 2. It can be seen that our approach achieves better performance than Rudnev's method (Rudnev et al., 2021) in MPJPE. Rudnev's method is learning-based and does not require the knowledge of the initial pose parameters. We use the network trained by Rudnev et al. (2021) which is limited to the resolution (240×180) of the DAVIS 240C camera. Thus, we simulate event streams of the same motion with the intrinsics provided in Rudnev et al. (2021) for Rudnev's method. We observe that since the global rotation and translation are fixed, events are only

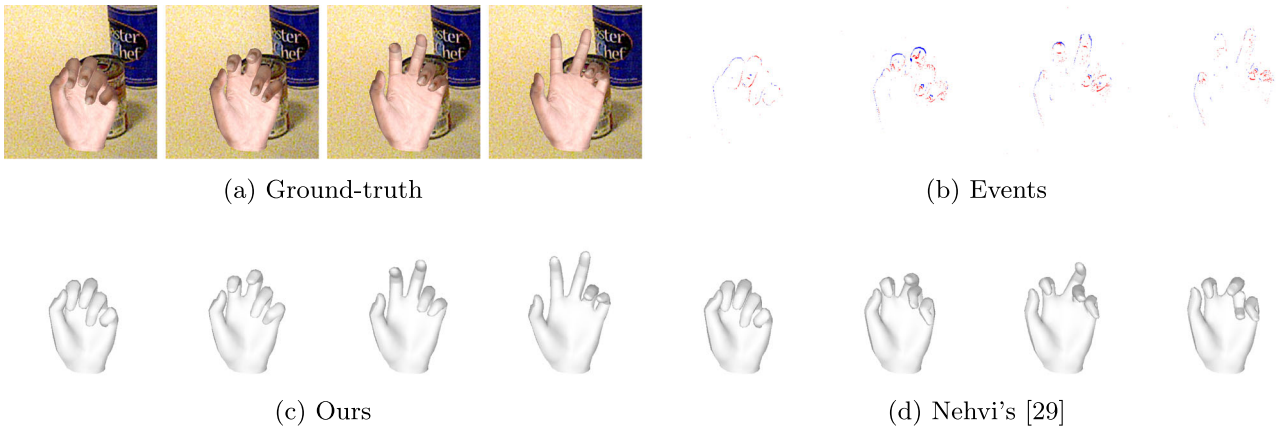
Table 2 Quantitative results on synthetic sequences

Scenario	Method	Mean MPJPE (mm)	Median MPJPE (mm)
MANO hand	Nehvi et al. (2021)	11.61	10.85
	Ours	**4.52**	**4.27**
SMPL-X hand	Rudnev et al. (2021)	11.88	10.73
	Ours	**1.11**	**0.76**
SMPL-X arm & Hand	Ours	**15.39**	**3.93**



(a) MANO hand sequences. (b) SMPL-X hand sequences. (c) SMPL-X arm+hand seqs.

Fig. 5 3D-PCK curve @50 mm on synthetic sequences



(a) Ground-truth (b) Events

(c) Ours (d) Nehvi's [29]

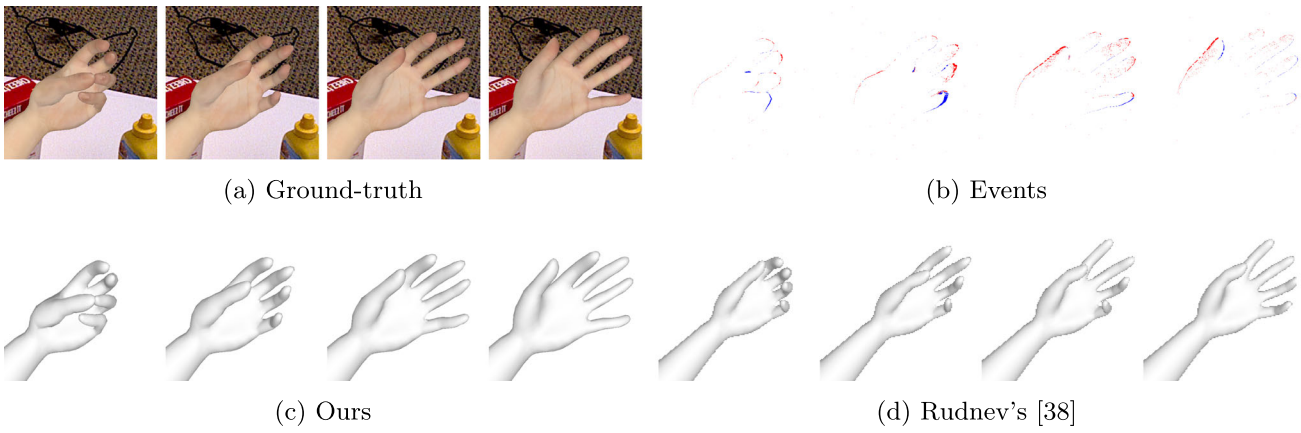
Fig. 6 Qualitative reconstruction results on synthetic MANO hand sequences

generated where the deformation occurs. Rudnev's method performs worse than our approach in this setting.

Finally, we evaluate the performance of our approach on sequences which combine arm and hand motion using the SMPL-X model. In the synthetic data generation process, we vary 6 principal parameters to synthesize hand poses and the 3 rotation parameters of the elbow joint. Our approach jointly optimize these hand and elbow parameters. The median MPJPE in Table 2 demonstrates that our approach can reconstruct the motion of the arm and hand with high accuracy. The mean MPJPE is higher than the median MPJPE due to failures in some sequences. We show failure cases and their analysis in the Sect. 6.9. The difference in accuracy to the SMPL-X hand sequences can be explained by the fact that for the SMPL-X arm & hand sequences, also the elbow joint needs to be reconstructed. Moreover, the hand is visible on different scales in the image (the hand is smaller for SMPL-X arm & hand). Hence, the absolute error in mm becomes higher.

As an incremental optimization-based approach, our approach can also drift, but it can snap the mesh silhouette to the observed events on the contour if sufficient observations are available. In Fig. 15, we provide a plot of median error over time for the MANO hand dataset.

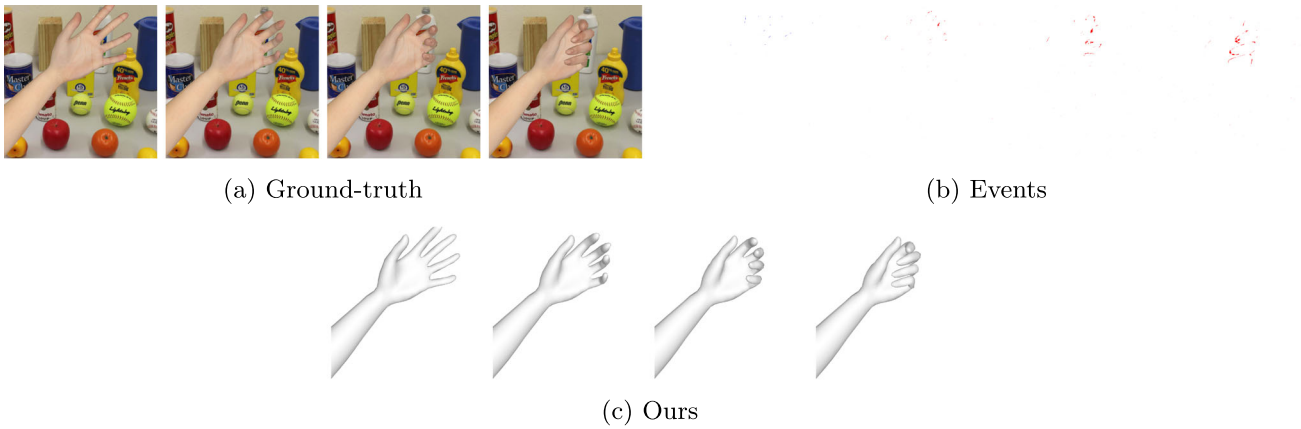
Due to the non-convexity of the problem, our approach needs a sufficiently good initial guess of the pose. In Sect. 6.6, we evaluate reconstruction accuracy vs. varying noise levels for the initial pose. We also evaluate the effect of varying noise in the events on accuracy.



(a) Ground-truth (b) Events

(c) Ours (d) Rudnev's [38]

Fig. 7 Qualitative reconstruction results on SMPL-X hand sequences



(a) Ground-truth (b) Events

(c) Ours

Fig. 8 Qualitative reconstruction results on SMPL-X hand sequences for an in-ward closing hand

6.5 Qualitative Evaluation

6.5.1 Synthetic Data

We show qualitative results of our approach and state-of-the-art baseline approaches (Nehvi et al., 2021; Rudnev et al., 2021) on synthetic sequences. For each object, the ground-truth RGB images, accumulated events during the motion, and reconstruction results are shown. We crop all images with a fixed ratio to increase the view of the objects. Results on synthetic MANO hand sequences of our approach and Nehvi's approach (Nehvi et al., 2021) are shown in Fig. 6c, d, respectively. It can be observed that our approach reconstructs the deformation of the hand well, while Nehvi's approach struggles to track the hand pose accurately. Note that Nehvi's method does assume black background and generatively models the specific log intensity changes induced at the optical flow at contours. Our approach only assumes that events are generated by contours without explicit dependency on the optical flow, hence, it is more robust to textured backgrounds. We compare our approach with Rudnev's method (Rudnev et al., 2021) in Fig. 7d on a SMPL-X hand deformation sequence. While our approach can reconstruct the hand motion well, Rudnev's approach performs less accurately. In Fig. 8 we show a qualitative example of a challenging SMPL-X hand motion in which an opened hand is bended inwards which causes self-occlusions at the palm and fingers. For the sequences with combined arm and hand motion of the SMPL-X model, we show qualitative results of our approach in Fig. 9c. Our proposed approach can reconstruct the motion well.

6.5.2 Real Data

In Fig. 10, we also show qualitative results of our approach with the MANO hand model on real sequences with hand motion captured with a DAVIS240C camera. The camera also records grayscale intensity frames for reference. Since our approach requires an initialization of the hand pose parameters, we use (Lin et al., 2021) on the first image frame and set rotation and translation manually, since the pretrained model did not yield proper poses on the DAVIS gray scale images.

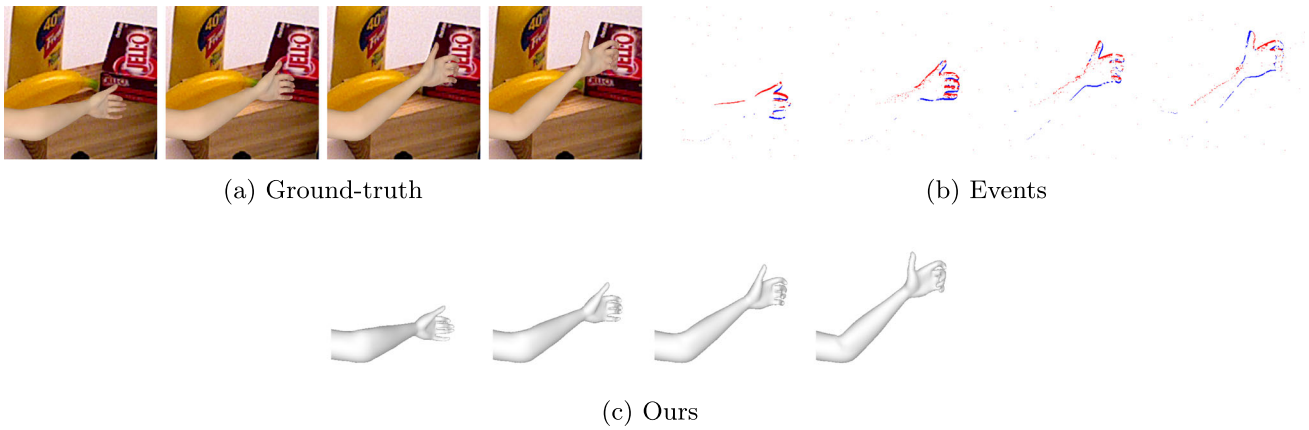


Fig. 9 Qualitative reconstruction results on SMPL-X arm & hand sequences

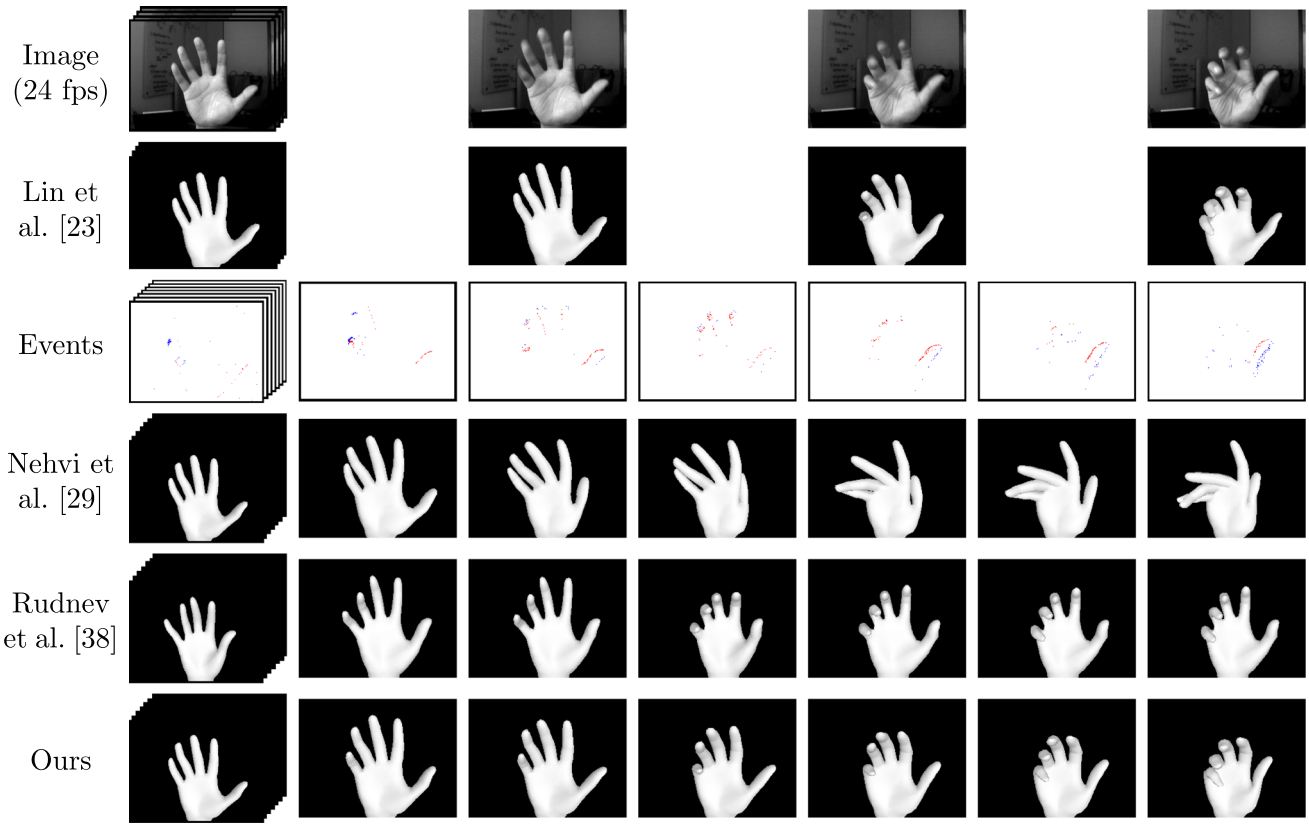


Fig. 10 Qualitative results on a real event sequence from a DAVIS240C camera. Lin et al. (2021) infer MANO pose parameters from intensity images; Rudnev et al. (2021) infer 6 principal MANO pose parameters; Nehvi et al. (2021) and our approach optimize 45 MANO pose parameters. Our approach recovers the deformation most similar to the ground truth

Further details on the initialization procedure are provided in Sect. 6.2. We compare our approach qualitatively with state-of-the-art image-based (MeshGraphormer (Lin et al., 2021)) and event-based (Nehvi et al., 2021; Rudnev et al., 2021) methods. MeshGraphormer is a learning-based approach which predicts MANO pose parameters from grayscale images. It has solid reconstruction performance for slower motions, but suffers from motion blur for fast motions. Furthermore, the temporal resolution of the reconstruction result is limited by the frequency of the frames. Compared to the result of MeshGraphormer (Lin et al., 2021) and event-based approaches (Nehvi et al., 2021; Rudnev et al., 2021), our approach follows the ground-truth reference more closely.

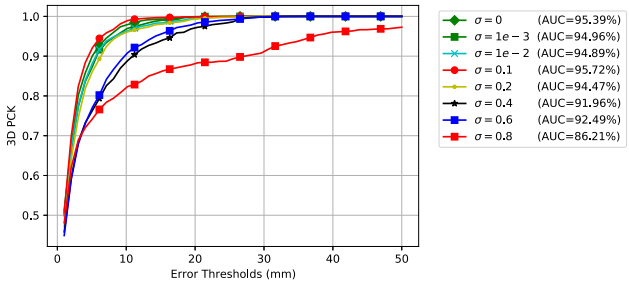


Fig. 11 Robustness to different levels of uncertainty for the initial parameters

6.6 Robustness

We evaluate robustness to noisy inputs on the SMPL-X hand motion sequences. In the first experiment, we investigate the robustness to noisy initial templates of objects. Here, we sample 6-dimensional initial pose parameters of hand model from a Gaussian distribution with the mean of ground-truth values and different standard deviations. The 3D-PCK curve and AUC value of each standard deviation are shown in Fig. 11. The result demonstrates that our approach still achieves an AUC of 0.86 when the standard deviation is 0.8. Note that a noise level of $\sigma = 0.2$ is already high for MANO hand parameters which are in the scale -2 to 2 (see Fig. 12 for hand parameters $\theta \in \mathbb{R}^6$).

In the second experiment, we evaluate robustness to noise in the input event stream. Noise is caused by the uncertainty of contrast threshold and salt-and-pepper noise in evaluation sequences. Here, we use different levels of standard deviation for contrast threshold sampling and vary the threshold for salt-and-pepper noise to simulate event streams with different noise levels for the same motion. We show the 3D-PCK curves and AUC values of different noise levels in Fig. 13.

The result in Fig. 13a demonstrates that our approach is robust to different levels of uncertainty for the contrast threshold in the event generation process. Besides, Fig. 13b shows that our approach has solid performance on different amounts of salt-and-pepper noise, too.

6.7 Ablation Study

6.7.1 Likelihood Formulation

In the first ablation study, we investigate variants of the data likelihood term formulated for E-step and M-step on SMPL-X hand motion sequences. The data likelihood for the E-step is formulated by the lateral probability, the longitudinal probability, and the contour probability:

$$p(x_i \mid z_i = j, \boldsymbol{\theta}) \propto p_{lateral} \cdot p_{longitudinal} \cdot p_{contour}. \quad (17)$$

Fig. 12 Variation of MANO hand parameters

(a) $\theta - 0.2 \cdot \mathbf{1}_6$ (b) $\theta \in \mathbb{R}^6$ (c) $\theta + 0.2 \cdot \mathbf{1}_6$

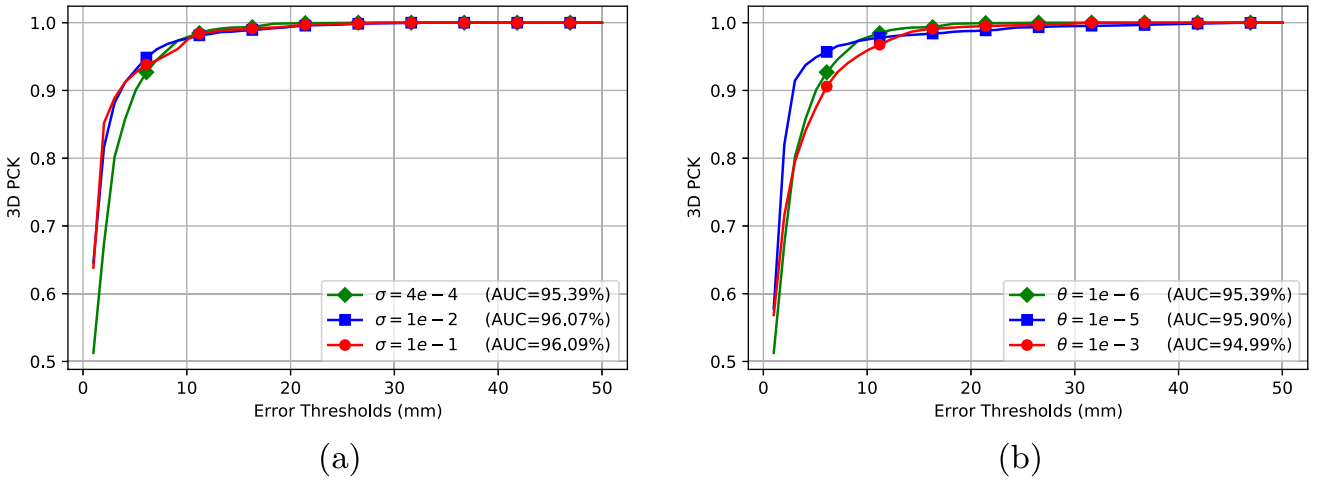


Fig. 13 Robustness to different level of **a** contrast threshold uncertainty; **b** salt-and-pepper noise

Table 3 Ablation study on probability terms of the data likelihood in the E-step and the M-step

$E_{lateral}$	$E_{longitudinal}$	E_{normal}	$M_{lateral}$	$M_{longitudinal}$	M_{normal}	MPJPE (mm)	AUC (%)
✓	✓	✓	✓	✗	✓	**1.5289**	**95.9308**
✓	✗	✓	✓	✗	✓	1.6523	93.5601
✓	✓	✗	✓	✗	✓	2.2500	92.7352
✓	✓	✓	✓	✗	✗	1.9891	92.8573

In the ablation study, we formulate the data likelihood in the E-step by either lateral probability and longitudinal probability in:

$$p(x_i \mid z_i = j, \boldsymbol{\theta}) \propto p_{lateral} \cdot p_{longitudinal}, \tag{18}$$

or the lateral probability and the contour probability:

$$p(x_i \mid z_i = j, \boldsymbol{\theta}) \propto p_{lateral} \cdot p_{contour}. \tag{19}$$

The proposed data likelihood in the M-step is formulated by the lateral probability and the longitudinal probability:

$$p(x_i \mid z_i = j, \boldsymbol{\theta}) \propto p_{lateral} \cdot p_{contour}. \tag{20}$$

In the ablation study, we formulate the data likelihood only with the lateral probability:

$$p(x_i \mid z_i = j, \boldsymbol{\theta}) \propto p_{lateral}. \tag{21}$$

We demonstrate the ablation study in the SMPL-X hand motion reconstruction. We combine different variants of E-step and M-step. The quantitative results of above mentioned variants are shown in Table 3.

The results demonstrate that the contour probability is essential for the formulation of the data likelihood term both in the E-step and the M-step. It also indicates that introducing longitudinal probability in the E-step can slightly improve the performance. Our full data likelihood formulation (Eqs. 17, 20) has best accuracy on the SMPL-X hand motion sequences.

6.7.2 Soft and Hard Association

In the second ablation study, we investigate the soft association and hard association in the M-step on SMPL-X hand motion sequences. For the soft association, we maximize the formulated objective for all mesh faces in the M-step. For the hard association, we select the mesh face which has the highest probability according to the E-step, and maximize only the likelihood for this mesh face in the M-step.

From the results in Table 4 it can be seen that the soft association is slightly better in MPJPE. The difference to hard association is small. We observe that the E-step often assigns a relatively high probability to one mesh face. Thus, the soft association and the hard association achieve similar results in this experiment.

Table 4 Ablation study on soft association and hard association

	MPJPE (mm)	AUC (%)
Soft association	1.11	96.38
Hard association	1.19	96.45

Table 5 Average event buffer processing run-time of Nevhi's approach and our method on the first MANO hand sequence

Method	Avg. runtime (s)
Nehvi's (Nehvi et al., 2021)	13.27
Ours	11.78

6.8 Run-Time Analysis

Differently to Lin et al. (2021), Rudnev et al. (2021), our implementation of soft association is not real-time capable, due to the complete evaluation of all event measurement likelihoods for all mesh faces in the soft E-step. Depending on the motion, a sequence can be split into 50 to 500 event buffers. For each event buffer, the current average run-time of the method is 8.76 s on MANO hand sequences, and 50.72 s on SMPL-X arm & hand sequences. EventHands (Rudnev et al., 2021), a real-time capable learning-based approach, has the fixed time span for event buffers buffers and thus it's not directly comparable. Here, we focus on comparing the runtime with the optimization-based method by Nehvi et al. (2021). We provide a per-buffer runtime comparison on the first MANO hand sequence (with average buffer temporal length $0.011s$) in Table 5. It demonstrates that our approach is faster than Nehvi's method in processing each event buffer while both are still far from real-time processing.

We also evaluate the run-time efficiency of a hard association approach (see Sect. 6.7.2) which performs an E-step only every 10th M-step. Since the hard association only considers one face in the M-step, it is more efficient and performs faster than the soft association.

In Fig. 14, we visualize results for the individual sequences and for different number of M-steps for each buffer on a subset of 10 SMPL-X hand sequences. Each buffer contains

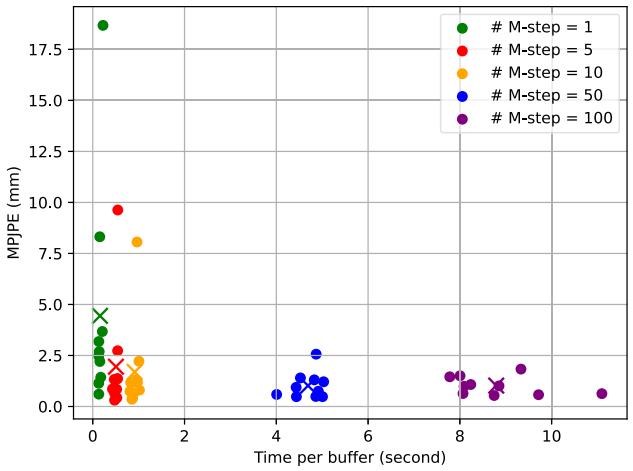


Fig. 14 Running time evaluation of SMPL-X hand sequences

Table 6 Run-time and accuracy on subset of SMPL-X hand sequences for hard association with different number of M-steps

	MPJPE (mm)	Time per buffer (s)
Soft association	1.36	43.57
1 M-step	4.44	0.15
5 M-steps	1.95	0.51
10 M-steps	1.69	0.91
50 M-steps	1.02	4.69
100 M-steps	1.02	8.79

200 events. We also show quantitative comparison between different number of M-steps in Table 6 over these sequences. We observe that with medium number of M-steps (e.g. 5–10), the accuracy of our approach degrades mildly. Note that these results cannot be directly compared to Table 4 because they are evaluated on a subset of the SMPL-X hand sequences and use a different learning rate of 0.007.

The soft association results on the same subset of sequences are shown in Table 6 as baseline. We can accelerate the run-time from 43.57 s per buffer with soft association 86.3 times for 5 M-steps per buffer and 47.6 times for 10 M-steps per buffer with hard association.

Although the hard association variants achieve better run-time results, they still do not reach real-time performance. Note that our current implementation is in PyTorch and that we use a first-order optimizer (ADAM (Kingma & Ba, 2015)). In future work, the optimization process could be implemented more efficiently by associating mesh faces with a local search, tailoring code with CUDA/C++, and using second-order Gauss-Newton methods instead of the current gradient-descent algorithm.

6.9 Drift and Failure Cases

6.9.1 Drift

As an incremental optimization-based approach, our approach can also drift, but it can snap the mesh silhouette to the observed events on the contour if sufficient observations are available. Figure 15 shows that for the MANO hand dataset our approach drifts from the ground-truth initial value at the beginning phase (buffers (0–100)), but is able to keep the same level of error in the remaining optimization process. The sequences in our experiments have a duration between 0.5 s and 2 s, while the number of buffers to optimize mainly depends on the speed of the motion.

6.9.2 Reconstruction Failures

Our approach fails in some sequences of SMPL-X body and hand motion. We visualize the ground-truth images, input event stream, and reconstructed arm and hand in Fig. 16. The initial pose is in the blue bounding box, and the final pose is in the green bounding box. Figure 16b shows that the hand at the initial pose does not generate valid events. The reason can be observed from Fig. 16a: the fingers at the initial pose have similar color as the background. According to the event generation model, no events are generated by the motion of fingers. The lack of events leads to a failure case of our approach in this sequence. However, as illustrated in Fig. 16c, our approach can still reconstruct the arm motion, because the arm moves over a different texture and sufficient events for the arm motion are generated.

The failure cases due to similar background and object color are more pronounced for the SMPL-X arm & hand than for SMPL-X hand sequences, because the hand appears smaller in the image and can overlap with the region in the background with similar color more strongly than on the SMPL-X hand sequences. For example, see Fig. 16b, where a large part of the events on the hand are missing. In the SMPL-X hand sequences, the hand appears larger (for example Fig. 9b) and the events are more widely distributed in the image, such that often only parts of the hand are affected and the hand pose is better constrained.

6.10 Assumption and Limitation

Our approach uses a loose coupling of frames and events by initializing the optimization from the gray-scale frame. A possible direction of future work is to extend the method by feeding the frame-based information at a specific lower rate and use the events to estimate pose between frames in a tightly-coupled joint optimization framework. In our experiments, self-occlusions occur within the hand (for instance between fingers, or fingers and the palm, see also Fig. 6). The

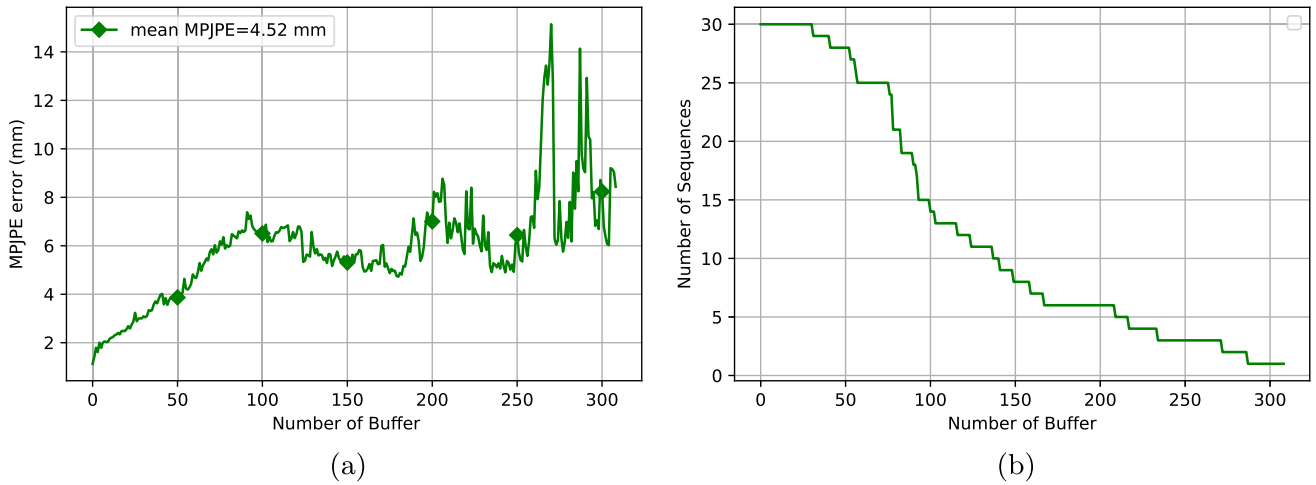


Fig. 15 Drift during optimization of MANO hand reconstruction experiments. **a** Average MPJPE development with the number of processed buffers. **b** Number of sequences still available at the number of processed buffers, indicating over how many sequences the MPJPE in (**a**) is averaged

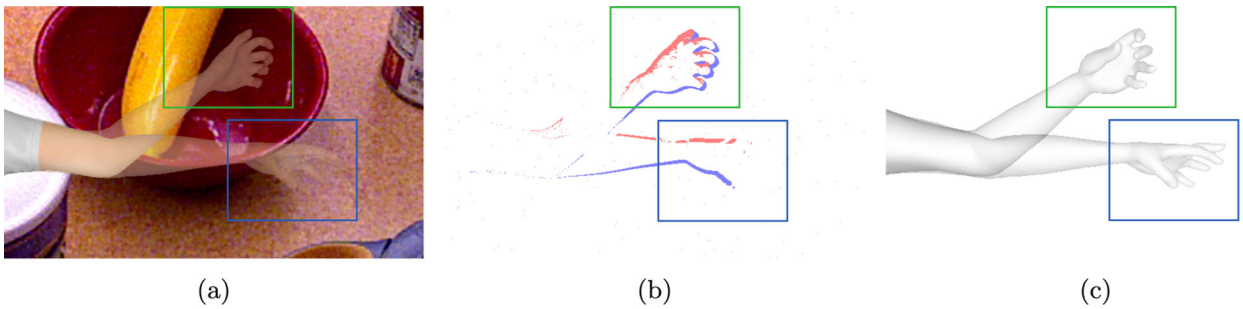


Fig. 16 Analysis of a failure case of our approach for SMPL-X arm & hand sequences. **a** Ground-truth motion in sequence 1. **b** Events in sequence 1. **c** Reconstructed arm and hand pose in sequence 1

more self occlusions, the more unconstrained the pose parameters get due to the partial observations and low number of events. The constant velocity model and the PCA subspace of MANO can help to regularize the motion in this partially constrained setting. Our method relies on events on the contour and cannot estimate deformation if there are little contour events due to similar background color or insufficient motion. We expect that for two independently moving hands that occlude each others, our method cannot track the motion of the occluded hand well due to the independence assumption of event measurements. Due to the image projection, the contour information seems not sufficient yet for reconstructing shape parameters concurrently with rotation and translation of the objects with our formulation. To address challenging settings like 6D pose estimation or crossing hands in future work, one could for instance investigate including learned temporal priors, texture-based cues, or combining events with frames in a joint optimization framework. Our approach assumes the deformation of non-rigid objects is constrained by a set of low rank parameters. Thus, adapting our approach to complex deformable models with higher degree of freedom could be challenging and can be considered as a future direction.

7 Conclusion

We present a novel non-rigid reconstruction approach for event cameras. Our approach formulates the reconstruction problem as an expectation-maximization problem. Events are associated to observed contours on parametrized mesh models and an alignment objective is maximized to fit the mesh parameters with event measurements. Our method outperforms qualitatively and quantitatively state-of-the-art event-based non-rigid reconstruction approaches (Nehvi et al., 2021; Rudnev et al., 2021). We also demonstrate that our proposed approach is robust to noisy events and initial parameter estimates. In future work, texture-based reconstruction from events and frames could be combined with our approach or the run-time of our implementation could be improved by searching for correspondences efficiently or using second-order optimization methods. In addition, our method could inspire novel learning-based approaches [e.g. (Messikommer

et al., 2020; Schaefer et al., 2022)], for instance, by using our EM objective to formulate self-supervised losses. Lastly, handling events in textured regions or tightly integrating our approach with image frame based measurements to further increase the variety of reconstructable objects could be an interesting direction for future research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-024-02011-z.

Acknowledgements This work was supported by Cyber Valley and the Max Planck Society. The authors thank director Bernhard Schölkopf of the Empirical Inference Department at the Max Planck Institute for Intelligent Systems for providing the DAVIS event camera. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Yuxuan Xue and Haolong Li.

Author Contributions All authors contributed to the design and development of the methodology. Implementation, data collection and experimental evaluation were performed by Yuxuan Xue. The first draft of the manuscript was written by Yuxuan Xue, Haolong Li, and Joerg Stueckler and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The datasets generated/analyzed during the experiments in this study cannot be made available due to license conditions of the SMPL, SMPL-X, and MANO models.

References

Aboukhadra, A. T., Malik, J., Elhayek, A., Robertini, N., & Stricker, D. (2023). Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 1001–1010).

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, G. Karypis (Eds.), *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019* (pp. 2623–2631). ACM.

Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., & Davis, J. (2005). Scape: Shape completion and animation of people. *ACM Transactions on Graphics (TOG), 24*(408–416), 07.

Bardow, P., Davison, A. J., & Leutenegger, S. (2016). Simultaneous optical flow and intensity estimation from an event camera. In *2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016* (pp. 884–892). IEEE Computer Society.

Bozic, A., Palafox, P. R., Zollhöfer, M., Dai, A., Thies, J., & Nießner M. (2020). Neural non-rigid tracking. In *Advances in neural information processing systems (NeurIPS)*.

Bregler, C., Hertzmann, A., & Biermann, H. (2000). Recovering non-rigid 3d shape from image streams. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 2690–2696).

Bryner, S., Gallego, G., Rebecq, H., & Scaramuzza, D. (2019). Event-based, direct camera tracking from a photometric 3d map using nonlinear optimization. In *International conference on robotics and automation, ICRA 2019, Montreal, QC, Canada, May 20–24, 2019* (pp. 325–331). IEEE.

Dai, Y., Li, H., & He, M. (2014). A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision, 107*(2), 101–122.

Davies, P. (1988). Kendall's advanced theory of statistics. In *Distribution theory* (Vol. 1).

Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., & Black, M. J. (2021). Collaborative regression of expressive bodies using moderation. In *2021 International conference on 3D vision (3DV 2021)* (pp. 792–804). IEEE.

Gallego, G., Rebecq, H., & Scaramuzza, D. (2018). A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018* (pp. 3867–3876). Computer Vision Foundation/IEEE Computer Society.

Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., & Scaramuzza, D. (2022). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 154–180.

Garg, R., Roussos, A., & Agapito, L. (2013). Dense variational reconstruction of non-rigid surfaces from monocular video. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1272–1279).

Gehrig, M., Millhäusler, M., Gehrig, D., & Scaramuzza, D. (2021) E-raft: Dense optical flow from event cameras. In *International conference on 3D vision (3DV)*.

Kim, H., Leutenegger, S., & Davison, A. J. (2016). Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European conference on computer vision (ECCV)* (pp. 349–364).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings*.

Lamarca, J., Parashar, S., Bartoli, A., & Montiel, J. M. M. (2021). Defslam: Tracking and mapping of deforming scenes from monocular sequences. *IEEE Transactions on Robotics, 37*(1), 291–303.

Li, H., & Stueckler, J. (2021). Tracking 6-dof object motion from events and frames. In *IEEE international conference on robotics and automation (ICRA)* (pp. 14171–14177).

Li, J., Bian, S., Liu, Q., Tang, J., Wang, F., & Lu, C. (2023). NIKI: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Li, J. X., Chao, C., Zhicun, B., Siyuan, Y. L, & Lu, C. (2021). Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*.

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics, 36*(6), 194:1-194:17.

Lichtsteiner, P., Posch, C., & Delbruck, T. (2008). A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits, 43*(2), 566–576.

Lin, K., Wang, L., & Liu, Z. (2021). Mesh graphormer. In *2021 IEEE/CVF international conference on computer vision, ICCV 2021, Montreal, QC, Canada, October 10–17, 2021* (pp. 12919–12928). IEEE.

Liu, S., Chen, W., Li, T., & Li, H. (2019). Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *2019 IEEE/CVF international conference on computer vision (ICCV)* (pp. 7707–7716).

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions in Graphics (Proceedings of SIGGRAPH Asia), 34*(6), 248:1-248:16.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, 34*(6), 248:1-248:16.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017). Monocular 3d human pose estimation in the wild using improved CNN supervision. In *2017 Fifth International Conference on 3D vision (3DV)*. IEEE.

Messikommer, Ni., Gehrig, D., Loquercio, A., & Scaramuzza, D. (2020). Event-based asynchronous sparse convolutional networks. In *CoRR*. arXiv:2003.09148

Nehvi, J., Golyanik, V., Mueller, F., Seidel, H. P., Elgharib, M., & Theobalt, C. (2021). Differentiable event stream simulator for non-rigid 3d tracking. In *IEEE conference on computer vision and pattern recognition workshops* (pp. 1302–1311).

Newcombe, R. A., Fox, D., & Seitz, S. M. (2015). Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 343–352).

Ngo, D. T., Östlund, J., & Fua, P. (2016). Template-based monocular 3d shape recovery using Laplacian meshes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(1), 172–187.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., & Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *IEEE conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019* (pp. 10975–10985). Computer Vision Foundation/IEEE.

Rebecq, H., Gehrig, D., & Scaramuzza, D. (2018). ESIM: An open event camera simulator. In *2nd Annual conference on robot learning, CoRL 2018, Zürich, Switzerland, 29–31 October 2018, proceedings of machine learning research* (Vol. 87, pp. 969–982). PMLR.

Rebecq, H., Horstschaefer, T., & Scaramuzza, D. (2017). Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *British machine vision conference (BMVC)* .

Rebecq, H., Ranftl, R., Koltun, V., & Scaramuzza, D. (2019). Events-to-video: Bringing modern computer vision to event cameras. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 3852–3861). IEEE Computer Society.

Rebecq, H., Ranftl, R., Koltun, V., & Scaramuzza, D. (2021). High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(6), 1964–1980.

Romero, J., Tzionas, D., & Black, M. J. (2017). Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, 36*(6), 245:1-245:17.

Rudnev, V., Golyanik, V., Wang, J., Seidel, H. P., Mueller, F., Elgharib, M., & Theobalt, C. (2021). Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *IEEE/CVF international conference on computer vision (ICCV)* (pp. 12365–12375).

Salzmann, M., & Fua, P. (2009). Reconstructing sharply folding surfaces: A convex formulation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1054–1061).

Schaefer, S., Gehrig, D., & Scaramuzza, D. (2022). Aegnn: Asynchronous event-based graph neural networks. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12361–12371).

Sidhu, V., Tretschk, E., Golyanik, V., Agudo, A., & Theobalt, C. (2020). Neural dense non-rigid structure from motion with latent space constraints. *European Conference on Computer Vision (ECCV), 12361*, 204–222.

Stoffregen, T., & Kleeman, L. (2019). Event cameras, contrast maximization and reward functions: An analysis. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 12300–12308).

Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M. J., & Mei, T. (2021). Monocular, one-stage, regression of multiple 3d people. In *2021 IEEE/CVF international conference on computer vision, ICCV 2021, Montreal, QC, Canada, October 10–17, 2021* (pp. 11159–11168). IEEE.

Vidal, A. R., Rebecq, H., Horstschaefer, T., & Scaramuzza, D. (2018). Ultimate slam? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics and Automation Letters, 3*(2), 994–1001.

von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering accurate 3d human pose in the wild using IMUS and a moving camera. In *European conference on computer vision (ECCV)*.

Xiang, Y., Schmidt, T., Narayanan, V., & Fox, D. (2018). Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In H. Kress-Gazit, S. S. Srinivasa, T. Howard, & N. Atanasov (Eds.), *Robotics: Science and systems XIV* . Carnegie Mellon University.

Yu, R., Russell, C., Campbell, N. D. F., & Agapito, L. (2015). Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from RGB video. In *IEEE international conference on computer vision (ICCV)* (pp. 918–926).

Zhu, A. Z., Yuan, L., Chaney, K., & Daniilidis, K. (2018). Unsupervised event-based optical flow using motion compensation. In L. Leal-Taixé & S. Roth (Eds.), *Computer vision—ECCV 2018 workshops—Munich, Germany, September 8–14, 2018, proceedings, Part VI. Lecture notes in computer science* (Vol. 11134, pp. 711–714). Springer.