

Event-based Non-Rigid Reconstruction from Contours

Xuxuan Xue^{1,2}

yuxuan.xue@tuebingen.mpg.de

Haolong Li¹

haolong.li@tuebingen.mpg.de

Stefan Leutenegger²

stefan.leutenegger@tum.de

Joerg Stueckler¹

jorg.stueckler@tuebingen.mpg.de

¹ Max Planck Institute
for Intelligent Systems,
Tuebingen, Germany

² Technical University
of Munich,
Munich, Germany

Abstract

Visual reconstruction of fast non-rigid object deformations over time is a challenge for conventional frame-based cameras. In this paper, we propose a novel approach for reconstructing such deformations using measurements from event-based cameras. 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, we demonstrate the advantages of our method over state-of-the-art optimization and learning-based approaches for reconstructing the motion of human hands. A video of the experiments is available at <https://youtu.be/upcCVTomFXY>.

1 Introduction

Event cameras 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. 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. While several approaches for event-based cameras have been proposed for optical flow estimation or simultaneous localization and mapping [7], only little work has been devoted to non-rigid reconstruction [18, 76].

In this paper, we present a novel non-rigid reconstruction approach for event-based cameras. Our algorithm takes event streams 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). We propose a novel optimization-based method based

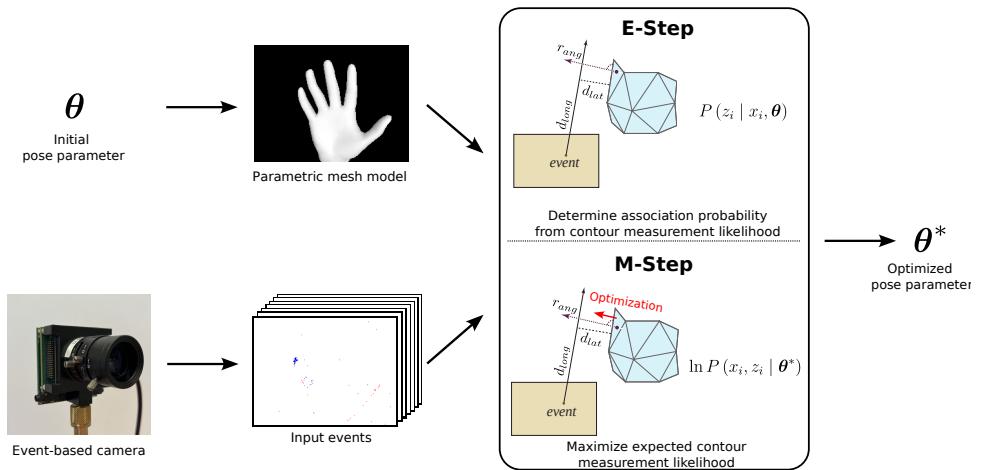


Figure 1: Our approach reconstructs non-rigid deformation states of objects from event streams of event-based cameras within an expectation maximization (EM) framework. In the E-step, the association probability of events to contour mesh faces is estimated from the contour measurement likelihood. In the M-step, the expected value of the measurement likelihood over the the association probability is maximized for pose parameter θ^* .

on expectation maximization (EM). Our method models event measurements at contours in a probabilistic way to estimate the association likelihood of events to mesh faces and maximize the measurement likelihood. The approach is evaluated on synthetic and real data sequences, and improvements over the state-of-the-art optimization and learning-based methods for hand reconstruction are demonstrated. For generating the sequences, we develop a novel event-based camera simulator for non-rigid deforming objects. Details of our proposed simulator and a comparison with the state-of-the-art event-based camera simulators [28, 24, 26] can be found in the supplementary material.

2 Related Work

A multitude of approaches has been proposed in recent years for scene reconstruction and rigid tracking using event-based cameras [9, 22, 30]. Only recently, approaches have been presented that track 3D object motion [10].

One can distinguish state-of-the-art approaches into approaches that accumulate disparity space images based on multi-view stereo [23], maximize contrast of reprojected events in a reference frame [6, 24], or apply a generative event measurement model derived from the brightness constancy assumption [9].

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. [20, 22, 23]) or approaches that use regularizing assumptions such as low-rank approximations to achieve non-rigid structure from motion (e.g. [3, 1, 8, 11, 28]). RGB-D cameras simplify the task due to the availability of dense depth for which several methods have been proposed recently (e.g. [2, 14]).

Non-rigid reconstruction and tracking with event-based cameras has only recently at-

tained attention in the computer vision community. Nehvi et al. [18] propose a differentiable event stream simulator by subtracting renderings of parametrized hand models. The paper demonstrates the use of the simulator for non-rigid motion tracking from event streams by optimization. Rudnev et al. [20] train a deep neural network on synthetic event streams to estimate the deformation of a MANO [25] hand model. To input the event data into the neural network, they propose to represent the data in local time windows. Different to these methods, we propose geometric contour alignment in a probabilistic optimization framework.

Some research in recent years has been devoted to event stream simulation [18, 24, 26]. We propose an event data simulator which generates synthetic events and other data modalities of human body motion, especially of hand deformation by simulating events with an adaptive sampling rate.

3 Event-based Non-Rigid Reconstruction from Contours

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

3.1 Expectation Maximization Framework

We formulate the 4D reconstruction problem as maximum-a-posteriori estimation of the model parameters θ given the event observations \mathbf{x} from the event camera

$$\theta^* = \arg \max_{\theta} \ln p(\mathbf{x} | \theta) + \ln p(\theta), \quad (1)$$

where $p(\theta)$ is a constant-velocity prior on θ . The analytical formulation of the likelihood is difficult because there is no observable relation between measurement \mathbf{x} and model parameters θ available. 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 each event corresponds to a point on the observed contour of the object. We introduce the latent variable $z_i = j$ which represents the association between the event x_i and a mesh face of the object with index j .

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

$$\theta^* = \arg \max_{\theta} \sum_{i=1}^N \ln \sum_{j=1}^F p(x_i, z_i = j | \theta) + \ln p(\theta), \quad (2)$$

where N is the number of aggregated events in an event buffer and F is the number of mesh faces. In practice, we aggregate a fixed number of events into an event buffer, assume that the event observations are independent from each other, and optimize over this event buffer. The optimization of the next event buffer is initialized with the parameters from the previous buffer. In the E-step, we update a probabilistic belief on the latent association variables using

a variational approximation given the current estimate of parameters $\bar{\theta}$ from the previous iteration,

$$q(z_i) \leftarrow \arg \max_{q(z_i)} \sum_{j=1}^F q(z_i = j) \ln \frac{p(x_i, z_i = j \mid \bar{\theta})}{q(z_i = j)}. \quad (3)$$

The optimal solution of this step is $q(z_i) = p(z_i \mid x_i, \bar{\theta})$. 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,

$$\theta \leftarrow \arg \max_{\theta} \sum_{i=1}^N \sum_{j=1}^F q(z_i = j) \ln p(x_i, z_i = j \mid \theta) + \ln p(\theta). \quad (4)$$

The posterior includes terms for the expected value of the measurement likelihood under the association probability distribution and a prior term on the parameters. In the following, we explain the concrete form of the EM steps in our approach in detail.

3.2 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 observed a few pixels off the actual event location. If fingers are bent in front of the palm, events generated on the contour of the finger might hit palm mesh faces, but miss mesh faces on the finger which generate 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 it causes the event. Inspired by SoftRasterizer [12], we formulate the contour measurement likelihood as

$$p(x_i \mid z_i = j, \theta) \propto \sigma \left(\delta_j^i \frac{d_{\text{lat}}^2(i, j)}{\alpha} \right) \exp \left(-\frac{d_{\text{long}}(i, j)}{\beta} \right) \exp \left(-\frac{r_{\text{ang}}(i, j)}{\gamma} \right), \quad (5)$$

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 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. 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.

3.3 E- and M-Steps

In the E-step, the association probability is calculated from the measurement likelihood,

$$p(z_i = j | x_i, \theta) = \frac{p(x_i | z_i = j, \theta)p(z_i = j | \theta)}{\sum_{j'} p(x_i | z_i = j', \theta)p(z_i = j' | \theta)} = \frac{p(x_i | z_i = j, \theta)}{\sum_{j'} p(x_i | z_i = j', \theta)}, \quad (6)$$

where we assume that the prior probability of the latent variable is uniform. For the M-step, we evaluate the measurement likelihood as

$$p(x_i, z_i = j | \theta) = p(x_i | z_i = j, \theta)p(z_i = j | \theta) \propto \sigma\left(\delta_j^i \frac{d_{\text{lat}}^2(i, j)}{\alpha}\right) \exp\left(-\frac{r_{\text{ang}}(i, j)}{\gamma}\right). \quad (7)$$

Here we do not include the longitudinal distance term. Ideally, this term should assign a constant probability to mesh faces on the same occlusion layer. Notably the term does not depend continuously on the shape parameters. 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. The angular error term in 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., $\ln p(\theta) = k \|\mathbf{v} - \mathbf{v}'\|_2^2$, $\mathbf{v} = \frac{\theta - \theta'}{\Delta t}$, where θ' 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 θ based on the current estimate of \mathbf{v} .

4 Experiments

We evaluate and demonstrate our event-based non-rigid reconstruction approach on synthetic and real sequences using MANO and SMPL-X object models and involving random motions and various background textures. We provide qualitative and quantitative results, comparing with state-of-the-art baselines. 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 given in the supplementary material. Please also refer to the supplemental video for qualitative results.

4.1 Experiment Setup

Implementation details To compensate between the accuracy and the efficiency, we accumulate events in event buffers and optimize a single set of shape parameters for the whole event buffer. Similar to [30], 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. Our simulator simulates the Prophesee camera. We use the pinhole camera model for the event camera and assume the camera intrinsics are calibrated. We have several hyperparameters in our framework. Please refer to the supplementary file for the tuning process of hyperparameters. Our algorithm optimizes for the pose parameters of the MANO hand model [2] or the SMPL body model [16, 21]. In case of the hand model, the pose parameters are in PCA space and the MANO modelling approach reconstructs the

vertex offsets, which are used together with the canonical pose vertices to generate the posed mesh. For the SMPL model, the pose parameters are the orientation parameters and Linear Blend Skinning (LBS) is used to recover the posed mesh.

Datasets We generate synthetic datasets of sequences with three types of different objects, namely the MANO [23] hand (Fig. 3a), the SMPL-X [16, 21] hand (Fig. 4a), and the combined SMPL-X [16, 21] arm and hand (5a). 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 and 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 a custom event simulator with adaptive sampling rate to generate the synthetic sequences for the different object models. A detailed description of the event simulator can be found in the supplementary material. We simulate a Prophesee camera with the image size of 1280×720 . The event contrast threshold is 0.5. For each sequence, the background image is randomly chosen from a texture-rich indoor scene in the YCB video dataset [32]. 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} . For further details on the noise generation process in our simulator, please refer to the supplementary material.

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 [31]), the percentage of correct 3D Joints (3D-PCK [12]), and the area under the PCK-curve (AUC [12]) with thresholds ranging from 0 to 50 mm. For hand sequences, we consider the 15 hand skeleton joints. For arm and hand sequences, only the forearm and the hand have motion. Thus, we consider one wrist joint and 15 hand joints.

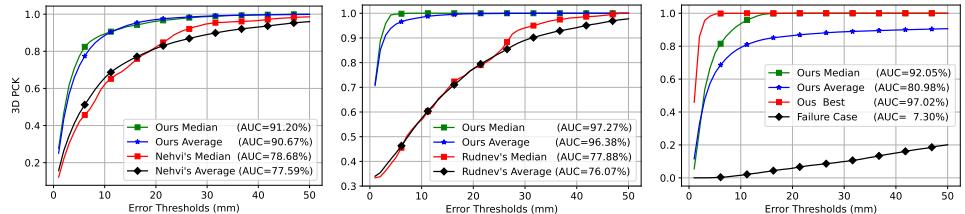
4.2 Quantitative Evaluation

We compare our approach quantitatively with the state-of-the-art event-based non-rigid object reconstruction methods: Nehvi’s optimization-based approach [18] evaluates using the MANO hand model, while Rudnev’s approach [26] 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 and 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 [23] hand model as the parametric mesh template. In 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 1. We also compare our method to the state-of-the-art event-based hand tracking approach [18] for the MANO model. 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 [11] to tune hyperparameters in Nehvi’s method and our method. We observe that our approach is about

Scenario	Method	mean MPJPE (mm)	median MPJPE (mm)
MANO hand	Nehvi et al. [18]	11.61	10.85
	Ours	4.52	4.27
SMPL-X hand	Rudnev et al. [26]	11.88	10.73
	Ours	1.11	0.76
SMPL-X arm & hand	Ours	15.39	3.93

Table 1: Results on synthetic sequences of different objects



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

Figure 2: 3D-PCK curve @50mm on synthetic sequences.

2.5-times more accurate than Nehvi’s method. We also show the 3D-PCK curve of both approaches in Fig. 2a. Apparently, our method has higher AUC than Nehvi’s method. Results in Table 1 and Fig. 2a demonstrate that our method outperforms Nehvi’s method clearly.

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 [26]. We report quantitative results in Table 1. It can be seen that our approach achieves better performance than Rudnev’s method [26] 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 [24] 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 [26] for Rudnev’s method. We observe that since the global rotation and translation are fixed, events are only generated where the deformation occurs. Rudnev’s method seems to perform less well in this case than our approach.

Finally, we evaluate the performance of our approach on sequences which combine arm

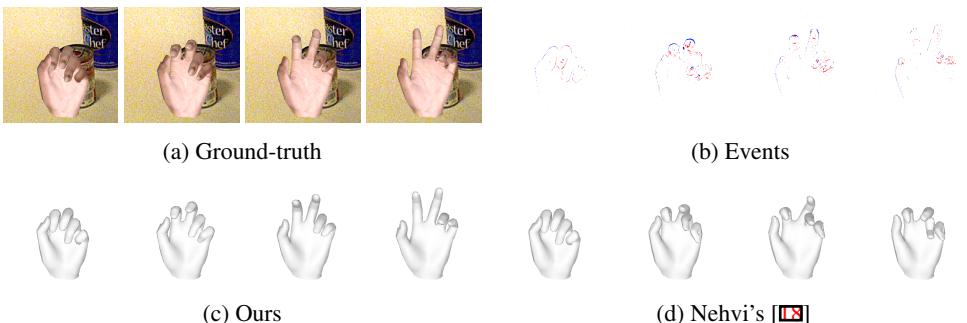


Figure 3: Qualitative reconstruction results on synthetic MANO hand sequences.

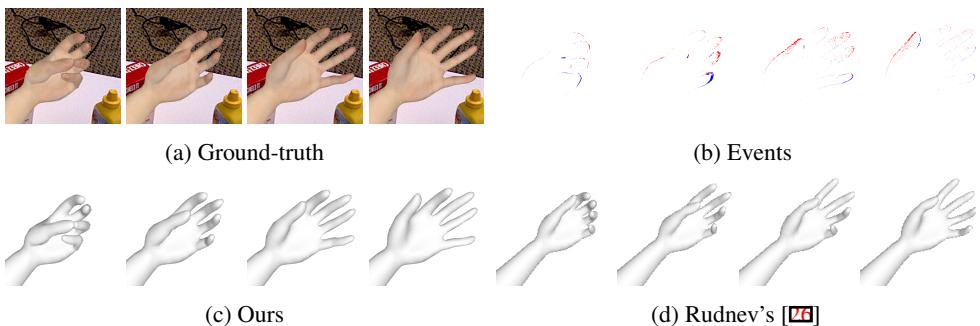


Figure 4: Qualitative reconstruction results on SMPL-X hand sequences.

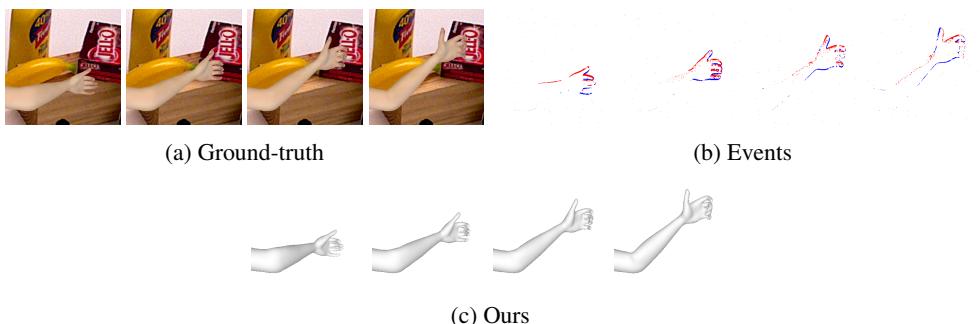


Figure 5: Qualitative reconstruction results on SMPL-X arm and hand sequences.

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 Tab. 1 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 supplementary material. 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 the supplementary material, we provide a plot of median error over time for the MANO hand dataset.

Differently to [13, 26], our implementation 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 seconds on MANO hand sequences, and 50.72 seconds on SMPL-X arm and hand sequences. Our current implementation in PyTorch. In future work, the optimization process could be implemented more efficiently by associating mesh faces with a local search, tailor code with CUDA/C++, and using second-order Gauss-Newton methods instead of the current gradient-descent algorithm. In

the supplementary material, we also provide evidence that hard data association with the most likely mesh face in the E-step only slightly affects accuracy. Due to the non-convexity of the problem, our approach needs a sufficiently good initial guess of the pose. In the supplementary material, 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 in the supplementary material.

4.3 Qualitative Evaluation

Synthetic Data We show qualitative results of our approach and state-of-the-art baseline approaches [18, 26] 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 [18] are shown in Figs. 3c and 3d, 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 Nehvis 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 [26] in Fig. 4d on a SMPL-X hand deformation sequence. While our approach can reconstruct the hand motion well, Rudnev’s approach performs less accurately. For the sequences with combined arm and hand motion of the SMPL-X model, we show qualitative results of our approach in Fig. 5c. Our proposed approach can reconstruct the motion well.

Real Data In Fig. 6, 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 [12] 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. Further details on the initialization procedure are provided in the supplementary material. We compare our approach qualitatively with state-of-the-art image-based (MeshGraphomer [1]) and event-based [18, 26] methods. MeshGraphomer 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 MeshGraphomer [1] and event-based approaches [18, 26], our approach follows the ground-truth reference more closely.

4.4 Assumptions and Limitations

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. 3). The 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

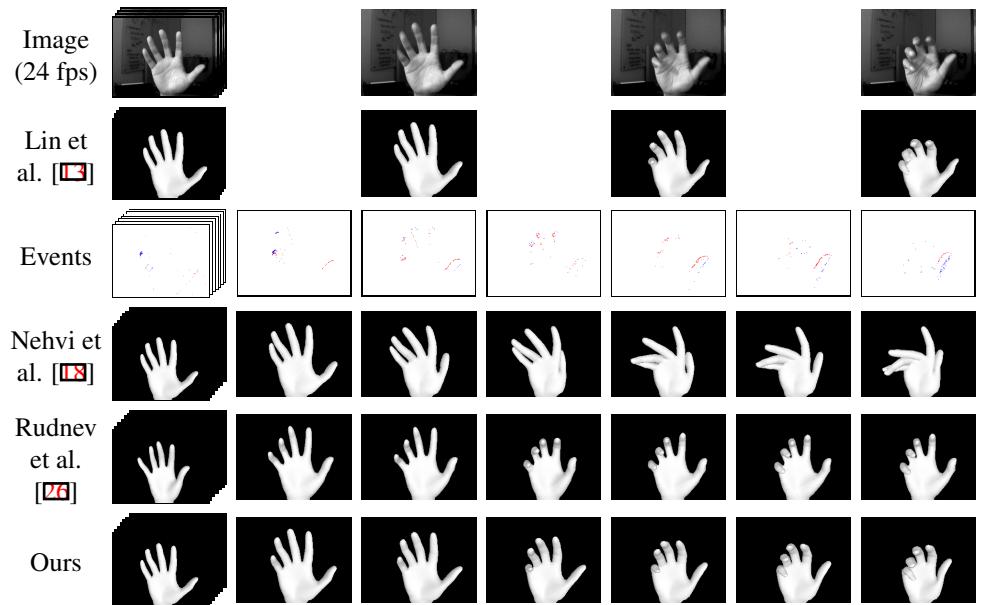


Figure 6: Qualitative results on a real event sequence from a DAVIS240C camera. Lin et al. [13] infer MANO pose parameters from intensity images; Rudnev et al. [26] infer 6 principal MANO pose parameters; Nehvi et al. [18] and our approach optimize 45 MANO pose parameters. Our approach recovers the deformation most similar to the ground truth.

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. 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.

5 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 [18, 26]. 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.

Acknowledgement

This work was supported by Cyber Valley and the Max Planck Society. The authors thank the Empirical Inference Department at Max Planck Institute for Intelligent Systems for providing the DAVIS event camera.

References

- [1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evinaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2623–2631. ACM, 2019. doi: 10.1145/3292500.3330701. URL <https://doi.org/10.1145/3292500.3330701>.
- [2] Aljaz Bozic, Pablo R. Palafox, Michael Zollhöfer, Angela Dai, Justus Thies, and Matthias Nießner. Neural non-rigid tracking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [3] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2690–2696, 2000.
- [4] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *Int. J. Comput. Vis.*, 107(2):101–122, 2014.
- [5] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV 2021)*, pages 792–804, Piscataway, NJ, December 2021. IEEE. doi: 10.1109/3DV53792.2021.00088.
- [6] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12280–12289, 2019.
- [7] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413. URL <https://doi.org/10.1109/TPAMI.2020.3008413>.
- [8] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1272–1279, 2013.
- [9] Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision (ECCV)*, pages 349–364, 2016.

- [10] Jose Lamarca, Shaifali Parashar, Adrien Bartoli, and J. M. M. Montiel. Defslam: Tracking and mapping of deforming scenes from monocular sequences. *IEEE Trans. Robotics*, 37(1):291–303, 2021.
- [11] Haolong Li and Joerg Stueckler. Tracking 6-dof object motion from events and frames. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 14171–14177, 2021.
- [12] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813. URL <https://doi.org/10.1145/3130800.3130813>.
- [13] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphomer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12919–12928. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01270. URL <https://doi.org/10.1109/ICCV48922.2021.01270>.
- [14] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7707–7716, 2019. doi: 10.1109/ICCV.2019.00780.
- [15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013. URL <https://doi.org/10.1145/2816795.2818013>.
- [16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.
- [17] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. doi: 10.1109/3dv.2017.00064. URL http://gvv.mpi-inf.mpg.de/3dhp_dataset.
- [18] Jalees Nehvi, Vladislav Golyanik, Franziska Mueller, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Differentiable event stream simulator for non-rigid 3d tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1302–1311, 2021.
- [19] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 343–352, 2015.
- [20] Dat Tien Ngo, Jonas Östlund, and Pascal Fua. Template-based monocular 3d shape recovery using laplacian meshes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):172–187, 2016.

- [21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 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*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01123. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Pavlakos_Expressive_Body_Capture_3D_Hands_Face_and_Body_From_a_CVPR_2019_paper.html.
- [22] Henri Rebecq, Timo Horstschafer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *British Machine Vision Conference (BMVC)*, 2017.
- [23] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. EMVS: event-based multi-view stereo - 3d reconstruction with an event camera in real-time. *Int. J. Comput. Vis.*, 126(12):1394–1414, 2018.
- [24] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, pages 969–982. PMLR, 2018. URL <http://proceedings.mlr.press/v87/rebecq18a.html>.
- [25] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883. URL <https://doi.org/10.1145/3130800.3130883>.
- [26] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgarib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12365–12375, 2021.
- [27] Mathieu Salzmann and Pascal Fua. Reconstructing sharply folding surfaces: A convex formulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1054–1061, 2009.
- [28] Vikramjit Sidhu, Edgar Tretschk, Vladislav Golyanik, Antonio Agudo, and Christian Theobalt. Neural dense non-rigid structure from motion with latent space constraints. In *European Conference on Computer Vision (ECCV)*, volume 12361, pages 204–222, 2020.
- [29] Timo Stoffregen and Lindsay Kleeman. Event cameras, contrast maximization and reward functions: An analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12300–12308, 2019.
- [30] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschafer, and Davide Scaramuzza. Ultimate slam? combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics Autom. Lett.*, 3(2):994–1001, 2018.

- [31] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018.
- [32] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov, editors, *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. doi: 10.15607/RSS.2018.XIV.019. URL <http://www.roboticsproceedings.org/rss14/p19.html>.
- [33] Rui Yu, Chris Russell, Neill D. F. Campbell, and Lourdes Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from RGB video. In *IEEE International Conference on Computer Vision (ICCV)*, pages 918–926, 2015.

A Event Simulation for Non-rigid Object

We propose an event data simulator which generates synthetic events and other data modalities (Fig. 7) of human body motion, especially of hand deformation. In addition to the events stream simulation, our simulator is able to simulate RGB image, depth map, 2D motion field, and normal map. Our event stream simulator is inspired by Nehvi's simulator [18], ESIM [24], and Rudnev's simulator [26]. It combines advantages of above mentioned simulators. We compare our simulator with these existing event stream simulators at the end of the section.

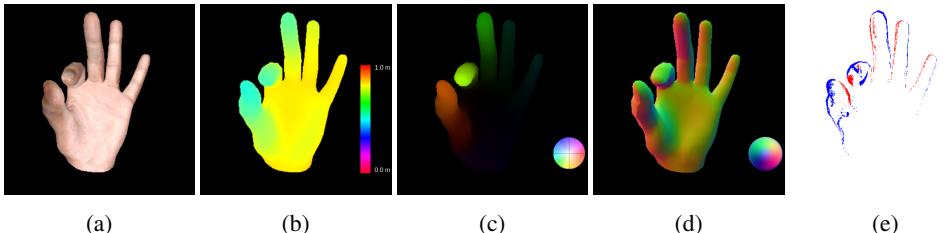


Figure 7: All data modalities in our simulator, including (a) RGB image, (b) depth map, (c) motion field, (d) normal map, (e) accumulated events in 1/30 seconds.

A.1 Event Generation Model

Unlike RGB cameras which capture absolute brightness for each pixels 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 with the logarithmic brightness value of the last sampled image at time t_{i-1} . The polarity p of the event is

$$p(\mathbf{u}_i, t_i) = \begin{cases} +1 & \text{if } \mathcal{L}(\mathbf{u}_i, t_i) - \mathcal{L}(\mathbf{u}_i, t_{i-1}) \geq C^+, \\ -1 & \text{if } \mathcal{L}(\mathbf{u}_i, t_{i-1}) - \mathcal{L}(\mathbf{u}_i, t_i) \geq C^-, \end{cases} \quad (8)$$

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 .

To simulate the motion field, we project the 3D movement of each mesh face onto the 2D image plane. Then, we adjust the time interval of next sample according to the largest motion vector magnitude among all pixels. For more details about the adaptive sampling principle, please refer to ESIM [24].

We also generate noisy events to make the simulated data more realistic. As in ESIM [24], 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 from a uniform distribution in $[0, 1]$, and compare 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 the adjustment of threshold to have the similar amount of salt-and-pepper noise as real event cameras, please refer to Rudnev's simulator [25].

A.2 Non-Rigid Parametric Models

When simulating hands alone using MANO [25], we use the full 45-dimensional PCA parameters of MANO. SMPL-X [26] is an expressive parametric human model, which models shape and pose of the human body using SMPL [25], hand pose using MANO [25], and facial expression using FLAME [8, 26]. Note that a PCA low-rank approximation of the pose parameters of the MANO model is used. The body pose is represented by 3-DoF orientations of 21 Joints while facial expression is controlled by 10 PCA parameters in expression space. We show the simulated data stream of these models in the supplementary file.

Our simulator takes a sequence of pose parameters of body and hand, the facial expression parameters as well as the simulation time as inputs and simulates event stream, RGB image, depth map, motion field and normal map (see Fig. 7). Similar as ESIM [24], our simulator assumes that the pose and expression parameters change linearly between two consecutive inputs of the sequence.

A.3 Comparison

We compare our proposed simulator with existing event stream simulators [23, 24, 26] in Table 2. Compared to Nehvi’s simulator [28], the simulator we propose can generate the 2D motion field for deforming objects and is accelerated using CUDA. Our experiments show that our simulator is 78-times faster by simulating the same hand motion of MANO model [25]. Compared to ESIM [24], our simulator can simulate events of human body motion. Compared to Rudnev’s simulator [26], our simulator uses the adaptive sampling strategy to avoid redundancy for small motion, while Rudnev’s method samples image frames every 0.001 seconds regardless of the motion.

	Objects	MF & AS	CUDA
Rebecq et al. [24]	Rigid Objects	✓	✗
Nehvi et al. [28]	MANO [25]	✗	✗
Rudnev et al. [26]	SMPL-H [25]	✗	✓
Ours	SMPL-X [26]	✓	✓

Table 2: Comparison between our event simulator and other event simulators. MF stands for motion field while AS stands for adaptive sampling.

MANO hand The simulated data stream of single MANO [25] model is shown in Fig. 8. Note that the MANO hand model only contains the hand but no arm.

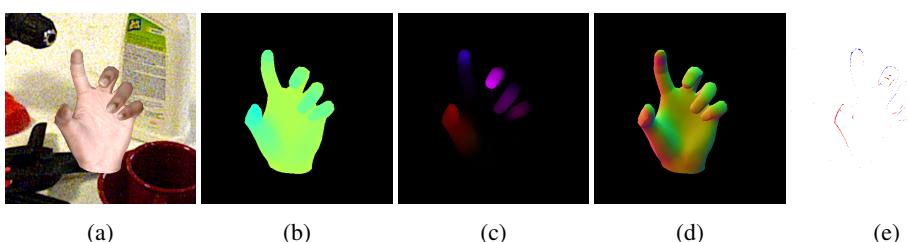


Figure 8: All data modalities of MANO [25] hand model, including (a) RGB image, (b) depth map, (c) motion field, (d) normal map, (e) accumulated events in 1/30 seconds.

SMPL-X hand We visualize the simulated data stream of SMPL-X [21] hand model in Fig. 9. It only contains the motion of the hand. However, attaching the arm to the hand makes it more realistic.

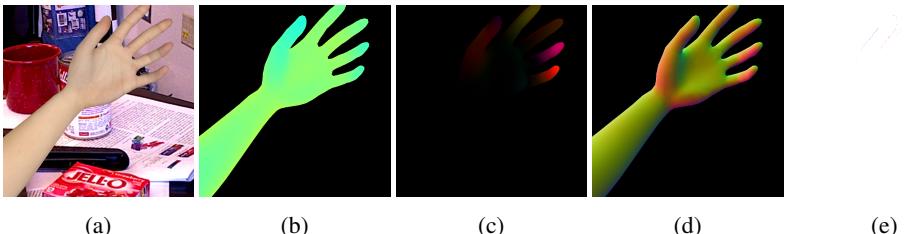


Figure 9: All data modalities of SMPL-X [21] hand, including (a) RGB image, (b) depth map, (c) motion field, (d) normal map, (e) accumulated events in 1/30 seconds.

SMPL-X arm and hand The simulated data modalities of SMPL-X [21] arm and hand motion are visualized in Fig. 10. The data contains the combined motion of the arm and the hand.

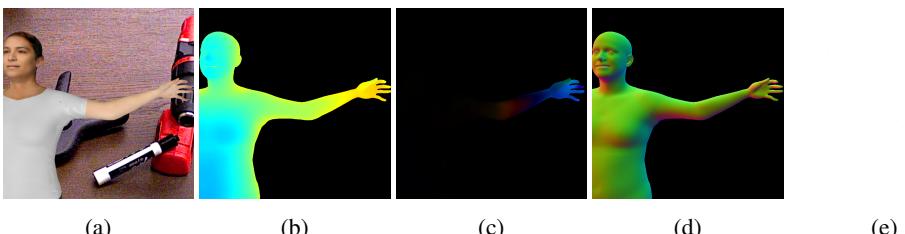


Figure 10: All data modalities in SMPL-X [21] arm & hand, including (a) RGB image, (b) depth map, (c) motion field, (d) normal map, (e) accumulated events in 1/30 seconds.

B Incremental EM reconstruction using Contour Events

B.1 Algorithm

In this section, we show the pseudo code of our approach in algorithm 1. It consists of an E-step to estimate the event association likelihood, and an M-step to maximize the association likelihood.

C Experiments

C.1 Real Data Mesh Template Initialization

As a template-based method, we assume the initial pose and shape parameters are known. For real data, we used MeshGraphomer [13] to infer MANO mesh model from a grayscale image. Then, we minimized the chamfer distance between the predicted hand mesh and the

Algorithm 1 Incremental EM reconstruction using event-based cameras

Input: events $\{e_k, \dots, e_{k+N-1}\}$ in spatio-temporal buffer W_k
Output: optimized mesh pose parameter θ_k

- 1: **procedure** EXPECTATIONMAXIMIZATION
- 2: $\theta_k \leftarrow$ initialization of mesh pose,
- 3: **E-step:**
- 4: $f(\theta_k) \leftarrow$ generate mesh model given pose parameter θ_k ,
- 5: $obj_func \leftarrow 0$, initialization of objective function
- 6: **for** e_i in $\{e_k, \dots, e_{k+N-1}\}$ **do**
- 7: $d_{normal}^i \in [F] \leftarrow$ dot product between event e_i to F faces,
- 8: $d_{lateral}^i \in [F] \leftarrow$ lateral distance from event e_i to F faces,
- 9: $d_{longitudinal}^i \in [F] \leftarrow$ longitudinal distance from event e_i to F faces,
- 10: $P(e_i|a, f(\theta)) \in [F] \leftarrow$ Likelihood of event e_i caused by F faces,
- 11: $E(LL(f(\theta_k|e_i, a))) \leftarrow$ expectaion of log-likelihood of event e_i ,
- 12: $obj_func \leftarrow obj_func + E(LL(f(\theta_k|e_i, a)))$,
- 13: **M-step:**
- 14: $\theta_k \leftarrow \underset{\theta_k}{\operatorname{argmax}} obj_func$
- 15: **if** Optimization not converged **then**
- 16: **goto** E-step.

PCA-parametrized MANO hand mesh to optimize shape and pose parameters of the captured hand. Finally, we fix the shape and pose of the hand, manually fine-tune the global rotation and translation of the mesh model by the visual alignment between the rendered 2D hand image and captured hand image.

C.2 Hyperparameter Tuning

We use Optuna [10] to tune hyperparameters in our approach and Nehvi’s method [18]. 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 is 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 will find the smallest MPJPE error for the hyperparameters. Depending on the scenarios, we have different settings of hyperparameters for the motion reconstruction based on the MANO model and the SMPL-X model.

C.3 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 11 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 results are between 0.5s

and 2s, while the number of buffers to optimize mainly depends on the speed of the motion.

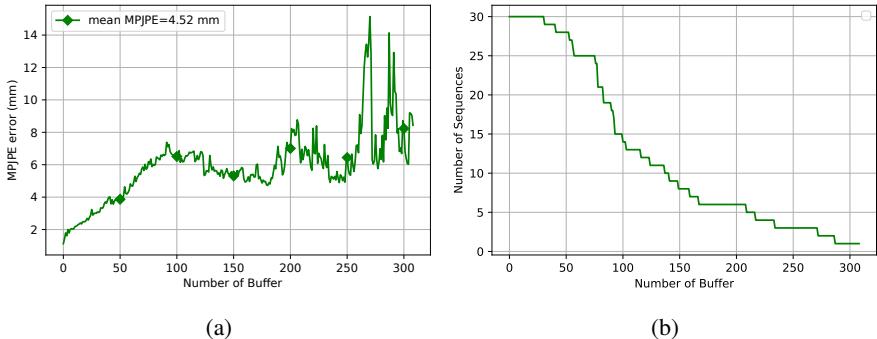


Figure 11: Drift during optimization of MANO hand reconstruction experiments. (a) Average MPJPE development with the number of processed buffer. (b) Number of sequences still available at the number of processed buffers, indicating over how many sequences the MPJPE in (a) is averaged.

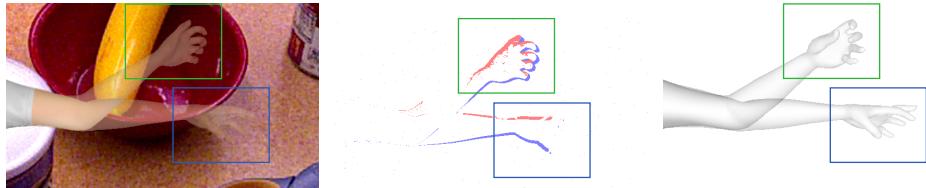
C.4 Failure Cases

Our approach fails in some sequences of SMPL-X body and hand motion. We investigated into why our approach fails in these cases. We visualize the ground-truth images, input event stream, and reconstructed arm and hand in figure 12. The initial pose is in the blue bounding box, and the final pose is in the green bounding box. Figure 12b shows that the hand at the initial pose does not generate valid events. The reason can be inferred from figure 12a: the fingers at the initial pose has the 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 the failure case of our approach in this sequence. However, as illustrated in figure 12c, our approach can still reconstruct the arm motion, because the events of arm motion are generated as usual.

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. 12b, 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. 10b) 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.

D Robustness to noise

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 in Fig. 13. The result demonstrates that our approach still has AUC of 0.86 when the standard deviation is 0.8. Note that a noise level



(a) Ground-truth motion in sequence 1.

(b) Events in sequence 1.

(c) Reconstructed arm and hand pose in sequence 1.

Figure 12: Analysis of the failure case of our approach for SMPL-X arm and hand sequences

of $\sigma = 0.2$ is already high for MANO hand parameters which are in the scale -2 to 2 (see figure 14 for hand parameters $\theta \in \mathbb{R}^6$).

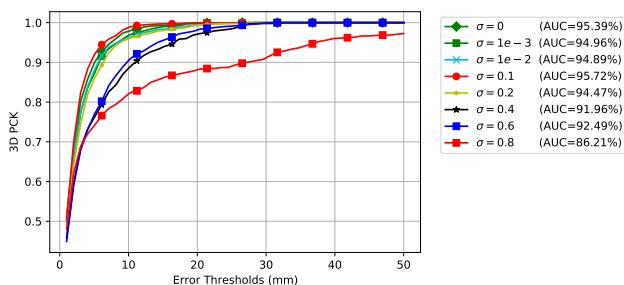


Figure 13: Robustness to different level of initial template noise

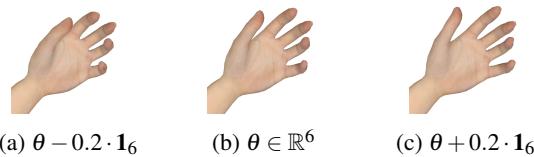
(a) $\theta - 0.2 \cdot \mathbf{1}_6$ (b) $\theta \in \mathbb{R}^6$ (c) $\theta + 0.2 \cdot \mathbf{1}_6$

Figure 14: Variation of MANO hand parameters

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 threshold for salt-and-pepper noise to simulate event streams of the different noise levels with the same motion. We show the 3D-PCK curves and AUC values of different noise levels in Fig. 15.

The result in Fig. 15a demonstrates that our approach is robust to different levels of uncertainty on contrast threshold in the event generation process. Besides, Fig. 15b shows that our approach has solid performance on different amounts of salt-and-pepper noise too.

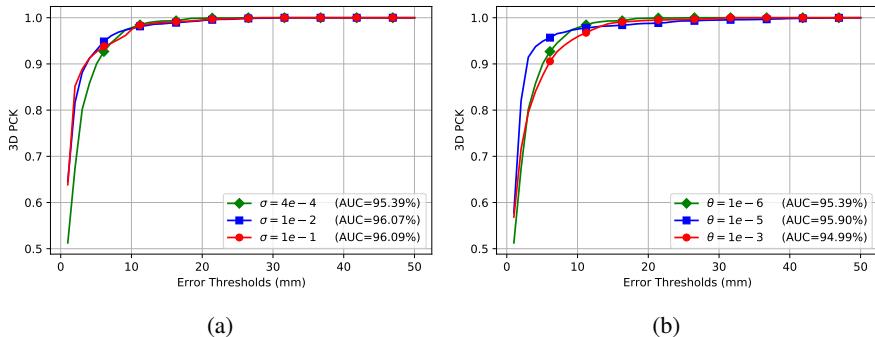


Figure 15: Robustness to different level of (a) contrast threshold uncertainty; (b) salt-and-pepper noise.

E Ablation Study

E.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 of E-step is formulated by the lateral probability, the longitudinal probability, and the contour probability:

$$P(x_i \mid z_i = j, \theta) \propto P_{lateral} \cdot P_{longitudinal} \cdot P_{contour}. \quad (9)$$

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

$$P(x_i \mid z_i = j, \theta) \propto P_{lateral} \cdot P_{longitudinal}, \quad (10)$$

or the lateral probability and the contour probability:

$$P(x_i \mid z_i = j, \theta) \propto P_{lateral} \cdot P_{contour}, \quad (11)$$

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

$$P(x_i \mid z_i \equiv j, \theta) \propto P_{lateral} \cdot P_{contour}, \quad (12)$$

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

$$P(x_i | z_i \equiv i, \theta) \propto P_{\text{latent}} \quad (13)$$

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

The quantitative results in table above 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 (Eq. 9, 12) has best accuracy on the SMPL-X hand motion sequences.

	MPJPE (mm)	AUC (%)
E3M2 (Eq. 9, 12)	1.5289	95.9308
E2 _{normal} M2 (Eq. 11, 12)	1.6523	93.5601
E2 _{longitudinal} M2 (Eq. 10, 12)	2.2500	92.7352
E3M1 _{lateral} (Eq. 9, 13)	1.9891	92.8573

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

E.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 logarithmic likelihood 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 the mesh face in the M-step.

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

Table 4: Ablation Study on soft association and hard association

The result in Table 4 shows that the soft association is slightly better in MPJPE. Generally, the soft and the hard association does not have huge difference. Our analysis is that the E-step already assigns a relatively high probability to one mesh face. Thus, the soft association and the hard association achieve similar results in this experiment.