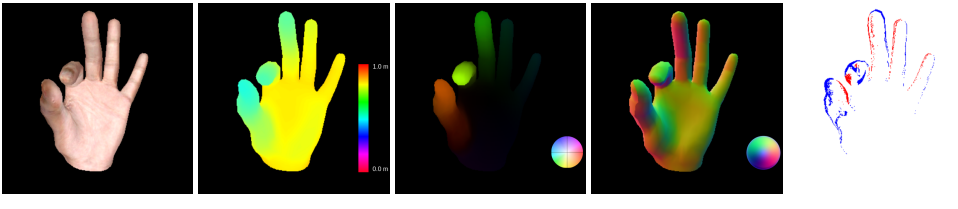
Supplementary Material – Event-based Non-Rigid Reconstruction from Contours

Yuxuan Xue[1,2]
yuxuan.xue@tuebingen.mpg.de
Haolong Li[1]
haolong.li@tuebingen.mpg.de
Stefan Leutenegger[2]
stefan.leutenegger@tum.de
Joerg Stueckler[1]
joerg.stueckler@tuebingen.mpg.de

[1] Max Planck Institute for Intelligent Systems, Tuebingen, Germany
[2] Technical University of Munich, Munich, Germany

1 Event Simulation for Non-rigid Object

We propose an event data simulator which generates synthetic events and other data modalities (Fig. 1) of human body motion, especially of hand deformation. In addition to the events stream simulation, our simulator is able to simulate RGB image, depth image, 2D motion field, and normal map. Our event stream simulator is inspired by Nehvi's simulator [6], ESIM [9], and Rudnev's simulator [10]. It combines advantages of above mentioned simulators. We compare our simulator with these existing event stream simulators at the end of the section.



(a) (b) (c) (d) (e)

Figure 1: 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.

1.1 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

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

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

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

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 [8].

We also generate noisy events to make the simulated data more realistic. As in ESIM [8], 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 [10].

1.2 Non-Rigid Parametric Models

When simulating hands alone using MANO [9], we use the full 45-dimensional PCA parameters of MANO. SMPL-X [7] is an expressive parametric human model, which models shape and pose of the human body using SMPL [5], hand pose using MANO [9], and facial expression using FLAME [2, 3]. 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. 1). Similar as ESIM [8], our simulator assumes that the pose and expression parameters change linearly between two consecutive inputs of the sequence.

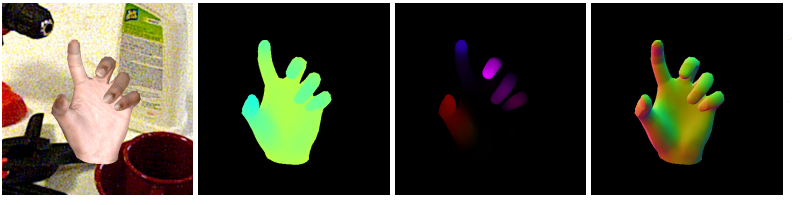
1.3 Comparison

We compare our proposed simulator with existing event stream simulators [6, 8, 10] in Table 1. Compared to Nehvi's simulator [6], 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 [9]. Compared to ESIM [8], our simulator can simulate events of human body motion. Compared to Rudnev's simulator [10], 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. [8]	Rigid Objects	✓	✗
Nehvi et al. [6]	MANO [9]	✗	✗
Rudnev et al. [10]	SMPL-H [9]	✗	✓
Ours	SMPL-X [7]	✓	✓

Table 1: Comparison between our event simulator and other event simulators. MF stands for motion field while AS stands for adaptive sampling.

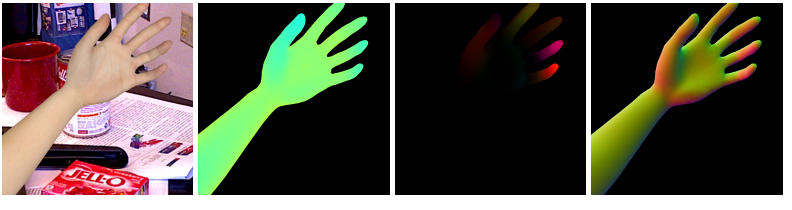
MANO hand The simulated data stream of single MANO [9] model is shown in Fig. 2. Note that the MANO hand model only contains the hand but no arm.



(a) (b) (c) (d) (e)

Figure 2: All data modalities of MANO [9] hand model, including (a) RGB image, (b) depth map, (c) motion field, (d) normal map, (e) accumulated events in 1/30 seconds.

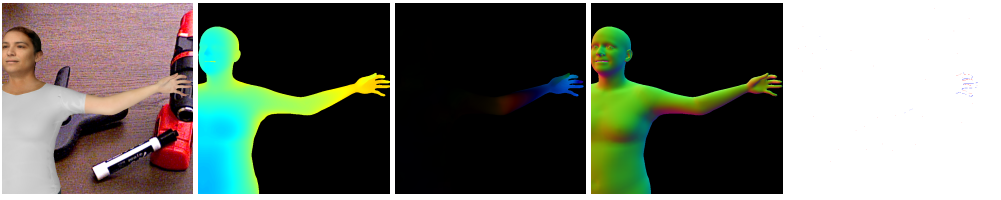
SMPL-X hand We visualize the simulated data stream of SMPL-X [7] hand model in Fig. 3. It only contains the motion of the hand. However, attaching the arm to the hand makes it more realistic.



(a) (b) (c) (d) (e)

Figure 3: All data modalities of SMPL-X [7] 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 [7] arm and hand motion are visualized in Fig. 4. The data contains the combined motion of the arm and the hand.



(a) (b) (c) (d) (e)

Figure 4: All data modalities in SMPL-X [7] arm & hand, including (a) RGB image, (b) depth map, (c) motion field, (d) normal map, (e) accumulated events in 1/30 seconds.

2 Incremental EM reconstruction using Contour Events

2.1 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.

Algorithm 1 Incremental EM reconstruction using event-based cameras

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

procedure EXPECTATIONMAXIMIZATION
  $\theta_k \leftarrow$ *initialization of mesh pose*,
  E-step:
  $f(\theta_k) \leftarrow$ *generate mesh model given pose parameter* θ_k,
  $obj_func \leftarrow 0$, *initialization of objective function*
  for e_i in $\{e_k, \cdots, e_{k+N-1}\}$ **do**
    $d^i_{normal} \in [F] \leftarrow$ *dot product between event* e_i *to* F *faces*,
    $d^i_{lateral} \in [F] \leftarrow$ *lateral distance from event* e_i *to* F *faces*,
    $d^i_{longitudinal} \in [F] \leftarrow$ *longitudinal distance from event* e_i *to* F *faces*,
    $P(e_i|a, f(\theta)) \in [F] \leftarrow$ *Likelihood of event* e_i *caused by* F *faces*,
    $E(LL(f(\theta_k|e_i, a))) \leftarrow$ *expectaion of log-likelihood of event* e_i,
    $obj_func \leftarrow obj_func + E(LL(f(\theta_k|e_i, a)))$,
  M-step:
  $\theta_k \leftarrow \underset{\theta_k}{\operatorname{argmax}}\ obj_func$
  if *Optimization not converged* **then**
    goto *E-step*.

3 Experiments

3.1 Real Data Mesh Template Initialization

As a template-based method, we assumes the initial pose and shape parameters are known. For real data, we used MeshGraphormer [4] to infer MANO mesh model from a grayscale

image. Then, we minimized 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 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.

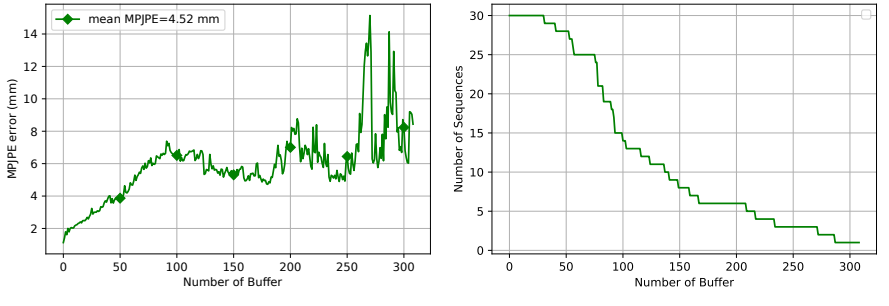
3.2 Hyperparameter Tuning

We use Optuna [1] to tune hyperparameters in our approach and Nehvi's method [6]. 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.

3.3 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 5 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.



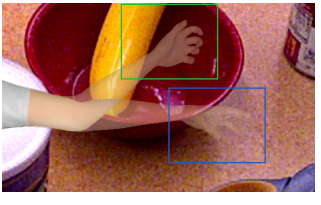
(a) (b)

Figure 5: 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.

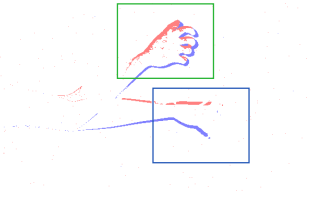
3.4 Failure Cases

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 groud-truth images, input event stream, and reconstructed arm and hand in figure 6. The initial pose is in the blue bounding box, and the final pose is in the green bounding box. Figure 6b shows that the hand at the initial pose does not generate valid events. The reason can be inferred from figure 6a: 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 6c, our approach can still reconstruct the arm motion, because the events of arm motion are generated as usual.

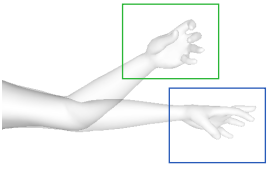
The failure cases due to similar background and object color are more pronounced for the SMPL-X arm & hand than for SMPL-X hand sequences, because the hand appears smaller in the image and can overlap with the region in the background with similar color more strongly than on the SMPL-X hand sequences. For example, see Fig. 6b, 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. 4b) 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.



(a) Ground-truth motion in sequence 1.



(b) Events in sequence 1.



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

Figure 6: Analysis of the failure case of our approach for SMPL-X arm and hand sequences

4 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. 7. The result demonstrates that our approach still has 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 figure 8 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 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. 9.

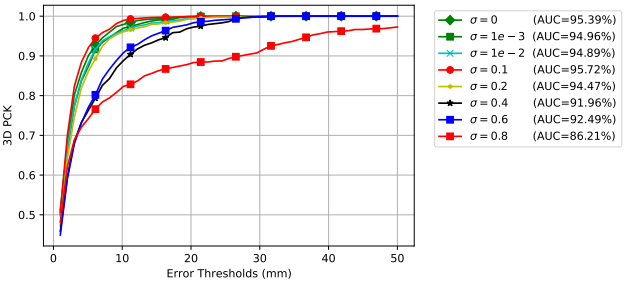


Figure 7: Robustness to different level of initial template noise

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

Figure 8: Variation of MANO hand parameters

The result in Fig. 9a demonstrates that our approach is robust to different levels of uncertainty on contrast threshold in the event generation process. Besides, Fig. 9b shows that our approach has solid performance on different amounts of salt-and-pepper noise too.

5 Ablation Study

5.1 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:

$$P(x_i \mid z_i = j, \theta) \propto P_{lateral} \cdot P_{longitudinal} \cdot P_{contour}. \tag{2}$$

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

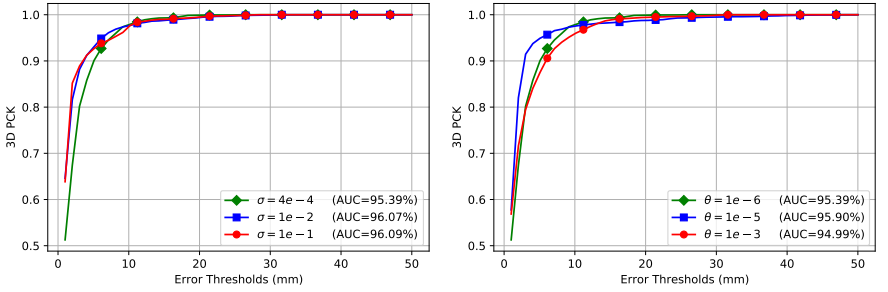
$$P(x_i \mid z_i = j, \theta) \propto P_{lateral} \cdot P_{longitudinal}, \tag{3}$$

or the lateral probability and the contour probability:

$$P(x_i \mid z_i = j, \theta) \propto P_{lateral} \cdot P_{contour}. \tag{4}$$

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

$$P(x_i \mid z_i = j, \theta) \propto P_{lateral} \cdot P_{contour}. \tag{5}$$



(a) (b)

Figure 9: Robustness to different level of (a) contrast threshold uncertainty; (b) salt-and-pepper noise.

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

$$P(x_i \mid z_i = j, \theta) \propto P_{lateral}. \tag{6}$$

We demonstrate the ablation study in the SMPL-X hand motion reconstruction. The quantitative results of above mentioned variants are shown in Table 2.

	MPJPE (mm)	AUC ($\%$)
E3M2 (Eq. 2, 5)	**1.5289**	**95.9308**
$\text{E2}_{\text{normal}}\text{M2}$ (Eq. 4, 5)	1.6523	93.5601
$\text{E2}_{\text{longitudinal}}\text{M2}$ (Eq. 3, 5)	2.2500	92.7352
$\text{E3M1}_{\text{lateral}}$ (Eq. 2, 6)	1.9891	92.8573

Table 2: Ablation Study on probability terms of the data likelihood in the E-step and the M-step

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. 2, 5) has best accuracy on the SMPL-X hand motion sequences.

5.2 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.

The result in Table 3 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.

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

Table 3: Ablation Study on soft association and hard association

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, Evimaria 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] 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.

[3] 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.

[4] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, 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.