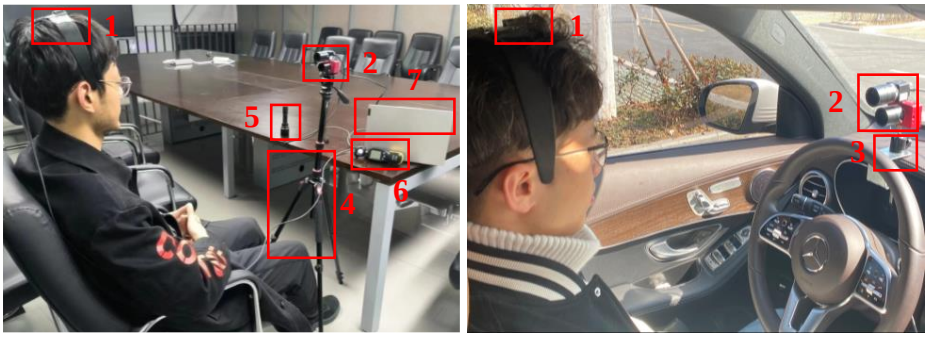
Supplementary Material: Event-based Head Pose Estimation: Benchmark and Method

Jiahui Yuan[1,*], Hebei Li[1,*], Yansong Peng[1], Jin Wang[1], Yuheng Jiang[1], Yueyi Zhang[1,†], and Xiaoyan Sun[1,2]

[1] University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{yuanjiahui,lihebei,pengyansong,jin01wang,jiangyuheng}@mail.ustc.edu.cn
{zhyuey,sunxiaoyan}@ustc.edu.cn



(a) (b)

Fig. 1: Visualization of the indoor(a) and in-car(b) scenarios from our data collection. (1) an IMU sensor; (2) the Prophesee Evk4 and Davis346 camera; (3) a vehicle-mounted metal bracket; (4) an adjustable tripod; (5) an artificial light source; (6) a photometer; (7) the laptop for data recording.

1 Dataset collection

Fig. 1 illustrates our data collection scenarios for indoor and vehicle environments. The primary difference lies in the support structures used: a tripod is employed indoors, while a metal mounting bracket is used within the vehicle. Notably, the remaining data collection equipment remains consistent across both settings. For clarity, only a representative subset of the equipment is depicted in the vehicle environment schematic.

The data collection process follows a specific order. First, we meticulously adjust the tripod (Indoor) or metal mounting bracket (In-car) to achieve optimal alignment with the volunteer's face. Next, the volunteer dons an IMU sensor to facilitate pose data collection. Subsequently, we utilize a photometer to assess the ambient lighting conditions. In scenarios requiring additional facial illumination, artificial lighting is strategically employed. Finally, the laptop triggers the camera

*Equal contribution †Corresponding author

and sensor programs, initiating the recording of both event streams and pose data.

2 Evaluation metrics

We compared our method with state-of-the-art HPE approaches, utilizing two metrics: Mean Absolute Error (MAE) and Mean Absolute Error Vector (MAEV). MAE is the standard metric for assessing the performance in HPE tasks. Assuming a set of ground truth Euler angles $\{P, Y, R\}$ and the predicted angles from the model as $\{\hat{P}, \hat{Y}, \hat{R}\}$, then MAE is defined as:

$$MAE = \frac{1}{3}(|P - \hat{P}| + |Y - \hat{Y}| + |R - \hat{R}|). \tag{1}$$

MAEV is a metric for evaluating the representation of rotation matrices. Assuming the reference rotation matrix representation is $R = [m_1, m_2, m_3]$ and the predicted rotation matrix from the model is $\hat{R} = [\hat{m}_1, \hat{m}_2, \hat{m}_3]$, then MAEV is defined as:

$$MAEV = \frac{1}{3}\sum_{i=1}^{3} \|m_i - \hat{m}_i\|_1. \tag{2}$$

3 More description for the loss calculation

In Sec. 4.4 of the main document, we introduced two functions $f_{GS}()$ and $f_{R\to E}()$ for loss calculation. Here we present description with more details about them.

f_{GS} represents the Gram-Schmidt mapping [**?**] [1]. Mathematically, it can be formulated as:

$$f_{GS}\left(\begin{bmatrix} | & | \\ p_1 & p_2 \\ | & | \end{bmatrix}\right) = \begin{bmatrix} | & | & | \\ q_1 & q_2 & q_3 \\ | & | & | \end{bmatrix}, \tag{3}$$

$$q_1 = \frac{p_1}{\| p_1 \|}, q_2 = \frac{s_2}{\| s_2 \|}, q_3 = q_1 \times q_2, \tag{4}$$

$$s_2 = p_2 - (q_1 \cdot p_2)q_1, \tag{5}$$

where the vectors p_1, p_2 are the retained column vectors of the matrix, while q_1, q_2, q_3 represent the column vectors after the mapping. The term s_2 denotes an intermediate variable, with the specific calculation process illustrated in the equations.

$f_{R\to E}(R)$ represents the conversion from the rotation matrix to Euler angles, which can be formulated as:

$$f_{R\to E}(R) = (x, y, z) = (Pitch, Yaw, Roll), \tag{6}$$

$$x = \arctan 2(R_{21}, R_{22}), z = \arctan 2(R_{10}, R_{00}), \tag{7}$$

$$y = \arctan 2(-R_{20}, sy), sy = \sqrt{R_{00}^2 + R_{10}^2}, \tag{8}$$

Table 1: Quantitative comparison of our method against other HPE methods with event-generated grayscale images as input on the Prophesee-HP dataset. The best results are highlighted in bold.

Method	Input	Euler angle errors				Vector errors			
	(10ms)	Pitch	Yaw	Roll	MAE	Left	Down	Front	MAEV
HopeNet [3]	Grayscale	11.51	11.90	7.30	10.24	15.67	14.76	18.23	16.22
FSA-Net [4]	Grayscale	12.60	13.21	6.82	10.88	16.57	15.51	20.25	17.44
WHENet [6]	Grayscale	11.21	11.00	7.59	9.93	15.23	14.70	17.32	15.75
6DRepNet [1]	Grayscale	8.69	9.02	5.64	7.78	12.01	11.08	13.59	12.23
TokenHPE [5]	Grayscale	14.04	14.64	6.89	11.86	17.60	17.13	22.81	19.18
Ours	Event	**6.00**	**7.47**	**4.55**	**6.01**	**9.89**	**8.20**	**10.49**	**9.53**

Table 2: Quantitative comparison of our method against other HPE methods with event-generated grayscale images as input on the Davis-HP dataset. The best results are highlighted in bold.

Method	Input	Euler angle errors				Vector errors			
	(10ms)	Pitch	Yaw	Roll	MAE	Left	Down	Front	MAEV
HopeNet [3]	Grayscale	11.94	13.18	6.58	10.57	16.03	14.92	19.87	16.94
FSA-Net [4]	Grayscale	12.45	13.08	6.28	10.60	15.83	15.17	20.17	17.06
WHENet [6]	Grayscale	11.63	12.23	6.12	9.99	14.89	14.29	18.90	16.03
6DRepNet [1]	Grayscale	9.98	10.72	5.80	8.83	13.43	12.47	16.21	14.04
TokenHPE [5]	Grayscale	13.21	12.75	5.98	10.65	15.33	15.66	20.47	17.15
Ours	Event	**7.50**	**7.57**	**4.41**	**6.49**	**9.75**	**9.32**	**11.65**	**10.24**

where R represents the rotation matrix, x, y, and z signify the Pitch, Yaw, and Roll angles, respectively. R_{ij} denotes the element in the i^{th} row and j^{th} column of the rotation matrix R. The term sy signifies the singularity value, while arctan2 represents the arctangent of the ratio of two values.

4 Comparison against methods with grayscale input generated from events

We present additional experiments comparing our event-based HPE method with existing methods that rely on event-generated grayscale images as input. For the existing methods, we utilized grayscale images reconstructed from event data on the Prophesee-HP and Davis-HP datasets using the E2VID method [2]. Our approach, however, continues to directly process the raw event streams, incorporating both temporal and spatial information.

Tab. 1 and Tab. 2 illustrate the performance comparison between our method and existing HPE methods on both datasets. It's important to note that existing HPE methods use single-channel grayscale images generated from 10ms event

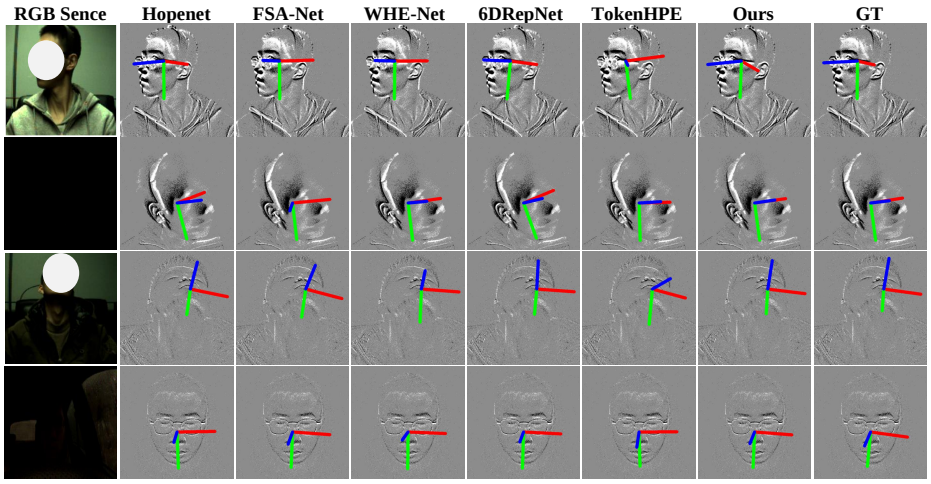


Fig. 2: Qualitative HPE results for four scene sequences using 10ms event slices on the Prophesee-HP dataset are demonstrated as follows: The leftmost column displays RGB scene illustrations, capturing the volunteer scenarios. The middle column provides a comparative analysis, contrasting the performance of existing Head Pose Estimation methods with our newly proposed approach. Lastly, the rightmost column showcases the actual angle labels.

slices as input. The results demonstrate that even though these grayscale-based methods share similarities with traditional HPE approaches, they are still outperformed by our method that leverages the richer information within the original event streams. This superiority can be attributed to the inevitable loss of information that occurs during the process of grayscale image reconstruction.

5 Additional qualitative results

We present additional qualitative results on the Prophesee-HP and Davis-HP datasets, as shown in Fig. 2 and Fig. 3, respectively. These figures showcase HPE estimations for a wider range of representative scenes within the datasets. It is evident from the visualizations that our method consistently outperforms other approaches by producing the most accurate estimations.

6 Event Representation

In our work, we convert asynchronous event streams from $t-1$ to t into corresponding voxel grids, denoted by $E_{t-1\rightarrow t} \in \mathbb{R}^{B\times H\times W}$, where B represents the number of temporal bins, and H and W represent the height and width of the grids [2] [7], respectively.

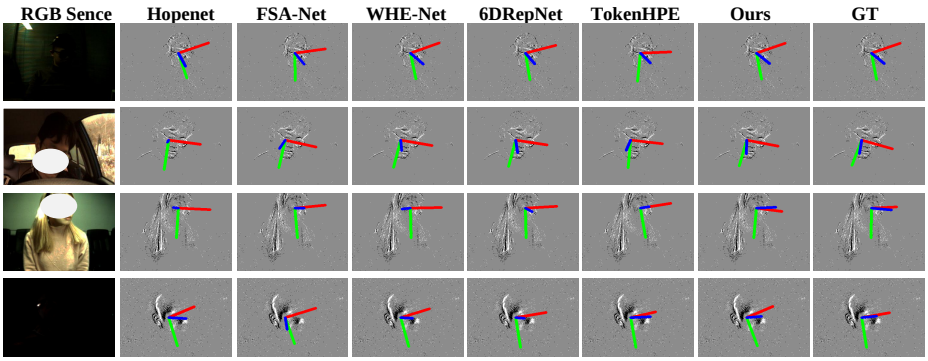


Fig. 3: Qualitative HPE results for four scene sequences using 10ms event slices on the Davis-HP dataset are demonstrated as follows: The leftmost column displays RGB scene illustrations, capturing the volunteer scenarios. The middle column provides a comparative analysis, contrasting the performance of existing Head Pose Estimation methods with our newly proposed approach. Lastly, the rightmost column showcases the actual angle labels.

$$V(x,y,t) = \sum_j [p_j \cdot \delta(x_j - x) \cdot \delta(y_j - y) \cdot \\ ((1-d_t) \cdot \delta(t_i - t) + d_t \cdot \delta(t_i + 1 - t))] \tag{9}$$

where δ is the Kronecker delta and $t_j^* = (B-1)\frac{t_j - t_0}{\Delta T}$, where B is the number of bins, ΔT is the time window of events, t_0 is the time of the first event in the window, $t_i = \lfloor t_j^* \rfloor$, and $d_t = t_j^* - t_i$.

References

1. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2496–2500. IEEE (2022) 2, 3
2. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3857–3866 (2019) 3, 4
3. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 2074–2083 (2018) 3
4. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1087–1096 (2019) 3
5. Zhang, C., Liu, H., Deng, Y., Xie, B., Li, Y.: Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8897–8906 (2023) 3

6. Zhou, Y., Gregson, J.: Whenet: Real-time fine-grained estimation for wide range head pose. arXiv preprint arXiv:2005.10353 (2020) 3
7. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019) 4