EyeGraph: Modularity-aware Spatio Temporal Graph Clustering for Continuous Event-based Eye Tracking Supplemental

Nuwan Bandara[†], **Thivya Kandappu**[†], **Argha Sen**[*], **Ila Gokarn**[‡], **Archan Misra**[†]
[†]Singapore Management University, [*]Indian Institute of Technology, Kharagpur
[‡]Singapore-MIT Alliance for Research and Technology (SMART)
{pmnsbandara, thivyak}@smu.edu.sg, arghasen10@gmail.com,
ila.gokarn@smart.mit.edu, archanm@smu.edu.sg

1 ML Reproducibility

1.1 Access to Dataset and Benchmark

The dataset and the implementation of our proposed benchmark can be found on our project webpage at `https://eye-tracking-for-physiological-sensing.github.io/eyegraph/`. Both the dataset and the source code are released under two licenses: (1) Creative Commons CC-BY-NC 4.0 license and (2) a custom license. The users/data requestors must agree to both licenses, and it is to be noted that if there is any conflict between any term(s) between two licenses, the custom license shall take priority over the Creative Commons CC-BY-NC 4.0 license.

1.2 Dataset Composition

As described in the corresponding main paper, *EyeGraph* Dataset is collected from 40 participants under three experimental setups: (1) conventional lab settings, (2) changing ambient illuminance and (3) user mobility, to capture event-based eye tracking in a wider range of practical and in-the-wild conditions.

Each session per participant in the conventional lab setting consists of four recordings, each lasting approximately four minutes. In the first two recordings, the participants wore the DAVIS346 camera whereas in the last two recordings, they wore the Pupil-Core eye tracker. The randomized movement pattern of the visual stimulus, i.e., the white circle was identical across the cross-device recording pair (for both the DAVIS346 and Pupil-Core device) but varied between the two recordings corresponding to the same wearable device. Therefore, under conventional lab settings, each participant has four recordings of near-eye tracking:

- Two recordings are event streams captured using DAVIS346, including:
  - event-based data
  - gray-scale image frames recorded at $30Hz$
  - inertial sensor measurements
  - external trigger data
- Two recordings are derived from the Pupil-Core eye tracking system, comprising:
  - raw near-eye videos at $\approx 100Hz$
  - point of gaze estimations at $\approx 100Hz$
  - pupil/iris segmentation
  - annotations for blinks and fixations

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

In the ambient luminance-changing settings, each session per participant (seated in an office environment similar to conventional lab settings) consists of four recordings. For the first two recordings, the participant wears the DAVIS346 sensor under two lighting conditions:

- *Constant Lighting Condition:* Near-eye Lux value maintained at 65 Lux throughout the experiment.
- *Variable Lighting Condition:* Near-eye Lux value alternates between 65 Lux and 8 Lux every 1-minute span.

For the last two recordings, the participant wears the Pupil-Core eye tracker under the two lighting conditions mentioned above.

The participant mobility settings mirror the ambient luminance-changing settings for data recording, however with two mobility conditions (with constant default lighting condition of near-eye 65 lux):

- *Stationary Condition:* Sitting in an office environment
- *Mobile Condition:* Moving freely while carrying a 14-inch laptop that displays the visual stimuli.

1.3 Model Hyperparameters

Here, we summarize the hyperparameters of our approach for easy reproducibility.

- Dynamic graph construction
  - We use min-max normalization on raw event data before constructing the graphs following [29]. Thus, λ_1 changes with the selected event volume while λ_2 and λ_3 change with the event data resolution (Section 5 in the corresponding paper).
  - We accumulate event volumes by setting the threshold for number of events $C = \{1500, 2000, 4000\}$ (Section 3.1).
  - Inspired by our empirical evaluations, unless stated otherwise, we set the max number of clusters to be considered as 5 due to the prominent anatomical clusters available in near-eye tracking: pupil, iris, lower and upper eyelids/lashes, and eyebrows as we observe in event data (Section 5 in the corresponding paper).
  - The scaling factor (λ) for statistical relevance in GMM fitting is set to be 0.1 while $N = 4$, $\xi_2 = 0.001$ and $\alpha = 1$ (Section 5 in the corresponding paper).
  - For Hawkes-based edge attribution, the number of divisions is set to be 8 while the decay factor is 0.5 and $\epsilon = 10^{-8}$ (Section 5 in the corresponding paper).
- Unsupervised topological clustering
  - With regard to model architecture, the latent vector is $64-$dimensional and the encoder has 8-layers (Section 4.1).
  - As described in Section 4.2, for all learning processes, the optimizer is set to be ADAM while the learning rate is constant and set to be 0.001. The batch size is 32 and the edges are split into training, validation, and test with proportions: 85%, 5%, and 10%. γ_1, γ_2 and γ_3 in the learning objective are 10, 10 and 0.1 respectively.

2 *EyeGraph* Dataset Details

In this section we explore the characteristics of our *EyeGraph* dataset.

2.1 Evaluation on Pupil and Gaze Distribution

2.1.1 Pupil Distribution

We empirically evaluate the pupil coordinates distribution, both spatially and temporally, using the Pupil-Core data to ensure that the collected data is instrumental, diverse and dense enough for multi-modal pupil tracking. The results from our investigation, which are illustrated in Figure 1 and 2, validate our above premise. Further, it shows that the collected data covers the spatial bounds while having a significant movement in temporal domain.

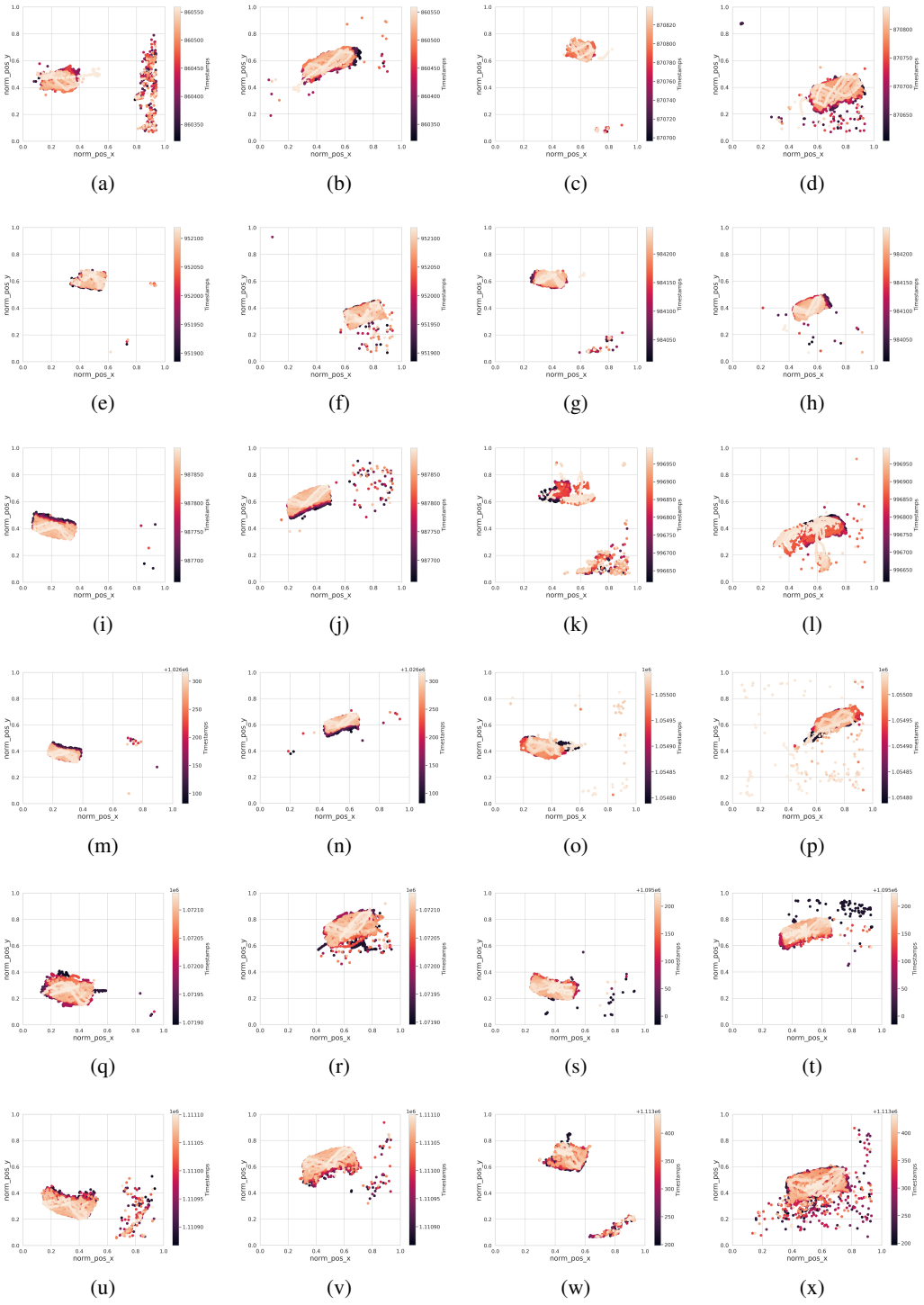
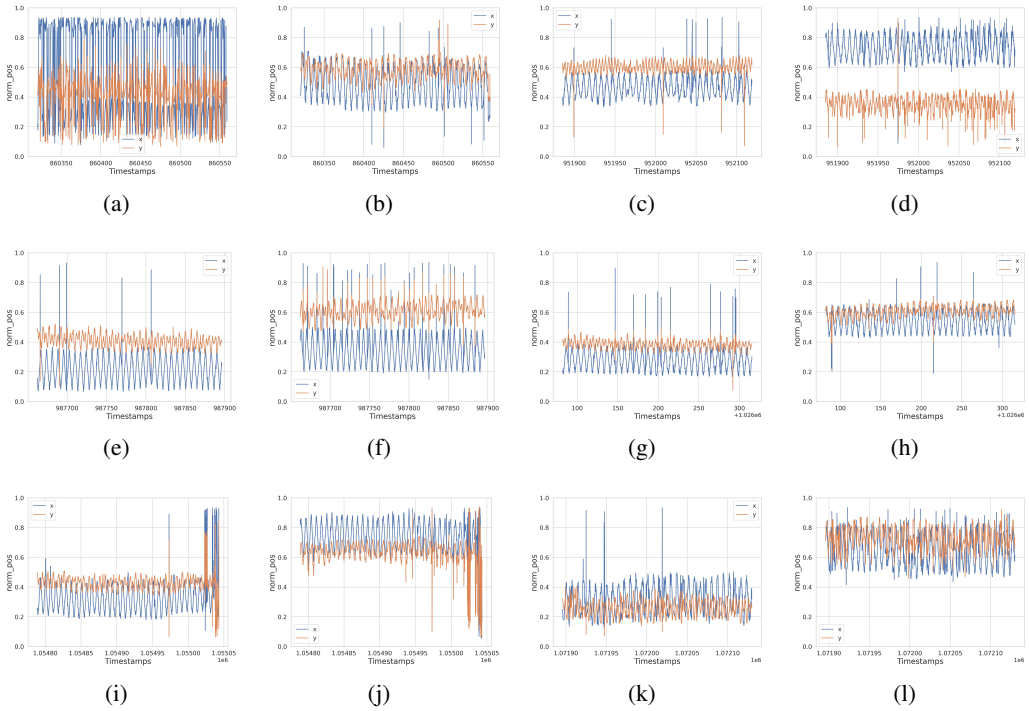


Figure 1: Pupil coordinates distribution (on XY plane) of randomly selected 12 participants with subject IDs: (a) 1 - right eye, (b) 1 - left eye, (c) 3 - right eye, (d) 3 - left eye, (e) 5 - right eye, (f) 5 - left eye, (g) 12 - right eye, (h) 12 - left eye, (i) 13 - right eye, (j) 13 - left eye, (k) 14 - right eye, (l) 14 - left eye, (m) 20 - right eye, (n) 20 - left eye, (o) 25 - right eye, (p) 25 - left eye, (q) 29 - right eye, (r) 29 - left eye, (s) 33 - right eye, (t) 33 - left eye, (u) 39 - right eye, (v) 39 - left eye, (w) 40 - right eye, (x) 40 - left eye; calculated from Pupil-Core data



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

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 2: Pupil coordinates propagation in temporal domain of randomly selected 6 participants with subject IDs: (a) 1 - right eye, (b) 1 - left eye, (c) 5 - right eye, (d) 5 - left eye, (e) 13 - right eye, (f) 13 - left eye, (g) 20 - right eye, (h) 20 - left eye, (i) 25 - right eye, (j) 25 - left eye, (k) 29 - right eye, (l) 29 - left eye; calculated from Pupil-Core data

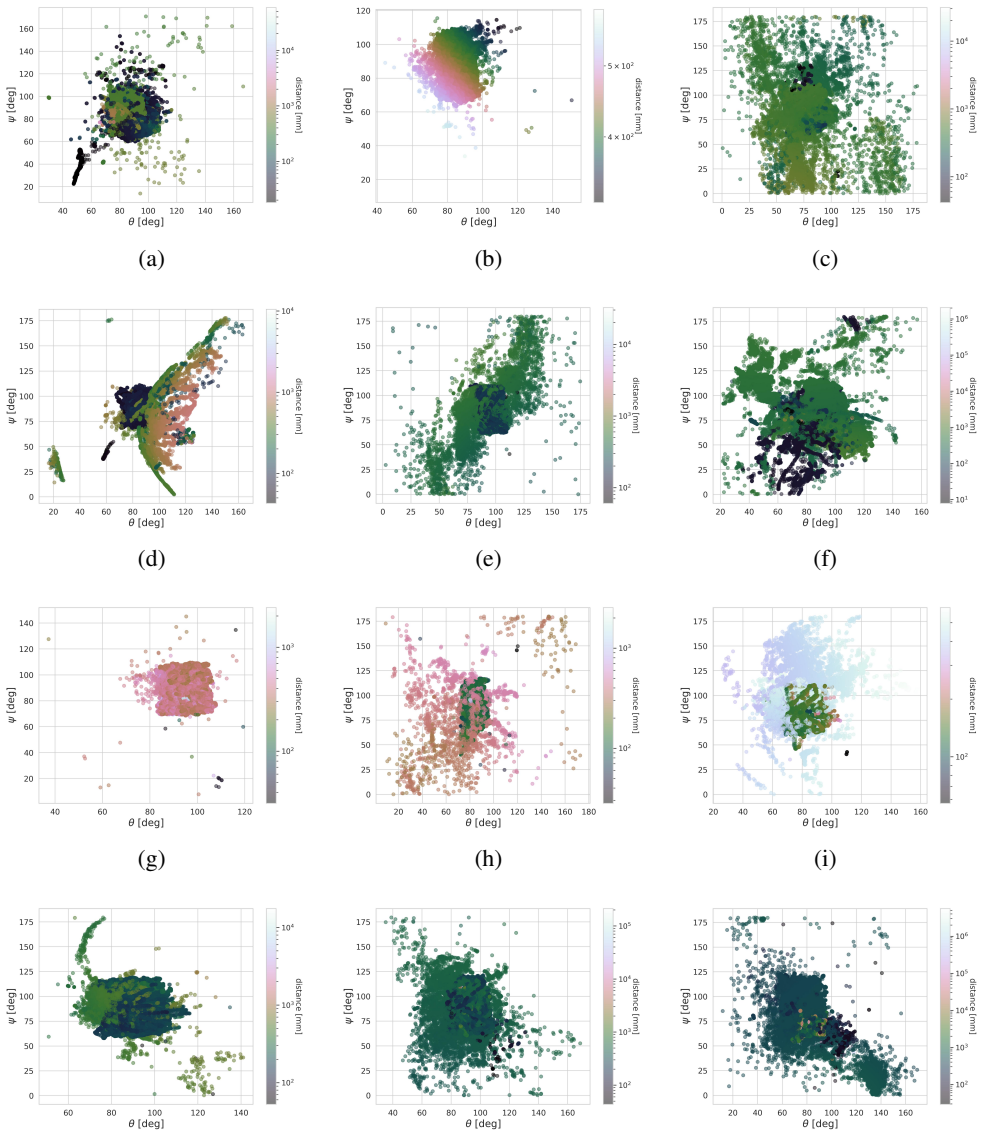
2.1.2 Gaze Distribution

To empirically evaluate the gaze movements each participant during the experiments, we calculate and visualize the gaze angles and velocities from the collected Pupil-Core data (for conventional lab settings) as depicted in Figure 3 and Figure 4 respectively. These figures emphasize the dense (while being inclusive of a wide range of eye movements) gaze distribution of the *EyeGraph* dataset with varying fields of view due to system mobility. This is a significant extension to the previously collected datasets [1, 41, 35] which are collected under fixed head postures. Further, it is evident that each participant exhibits distinctive gaze angle and velocity characteristics despite the same visual stimuli is followed by them.

2.2 Effect of Ambient Light Changes

We investigate the effect of ambient light changes for both event data and Pupil-Core data while focusing on its impact on the eye movements. In terms of Pupil-Core data, there are less notable differences in both pupil and gaze data since Pupil-Core primarily uses infrared-based method [16] in its estimation for those parameters. However, as depicted in Figures 5, 6 and 7, a key observation could be made upon the changes in pupil movement from default lighting to poor lighting condition: the spatial variation of pupil coordinates is slightly higher, thus more scattered in poor lighting condition than the default lighting condition, crucially due to (a) the changes in lighting leads the participant to momentarily loose the visual attention and thereby, the pupil location is changed, (b) the pupil dilation, and thereby the coordinates, is changed in different lighting conditions and (c) intrinsic noise in measurements is higher in poor lighting conditions.

As numerous studies revealed [9, 22], event cameras are considerably susceptible to being distorted from different kinds of noise including intrinsic (i.e., hardware), temporal, thermal and ambient light noise. As highlighted in [22], typically, the bright lighting leads to pixel saturation in event cameras and thereby diminishes the pixel sensitivity whereas the low lighting conditions causes to



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 3: Gaze angle distribution (in polar coordinates) of randomly selected 12 participants with subject IDs: (a) 1, (b) 3, (c) 5, (d) 12, (e) 13, (f) 14, (g) 20, (h) 25, (i) 29, (j) 33, (k) 39, (l) 40; calculated from Pupil-Core data

introduce significant spatial disparities by adding more measurement noise. To this end, here, we utilize three straight-forward filtering methods [14] to evaluate the noise levels in collected event data under default lighting condition and the poor lighting condition.

- Fixed-threshold local temporal neighbourhood filter (LTN): If an event is supported by another event on the same pixel within a set temporal threshold (for a given event volume), then the event is not considered as noise. Here, the temporal threshold is $100ms$ and the number of events for a given evaluation is 2000.

- Fast decay filter (FD): Here, the events are considered to be contributing to a low-resolution accumulated image and thus, an event is filtered if it is not supported by another event which is in the former event's spatio-temporal neighbourhood via a decaying temporal threshold.

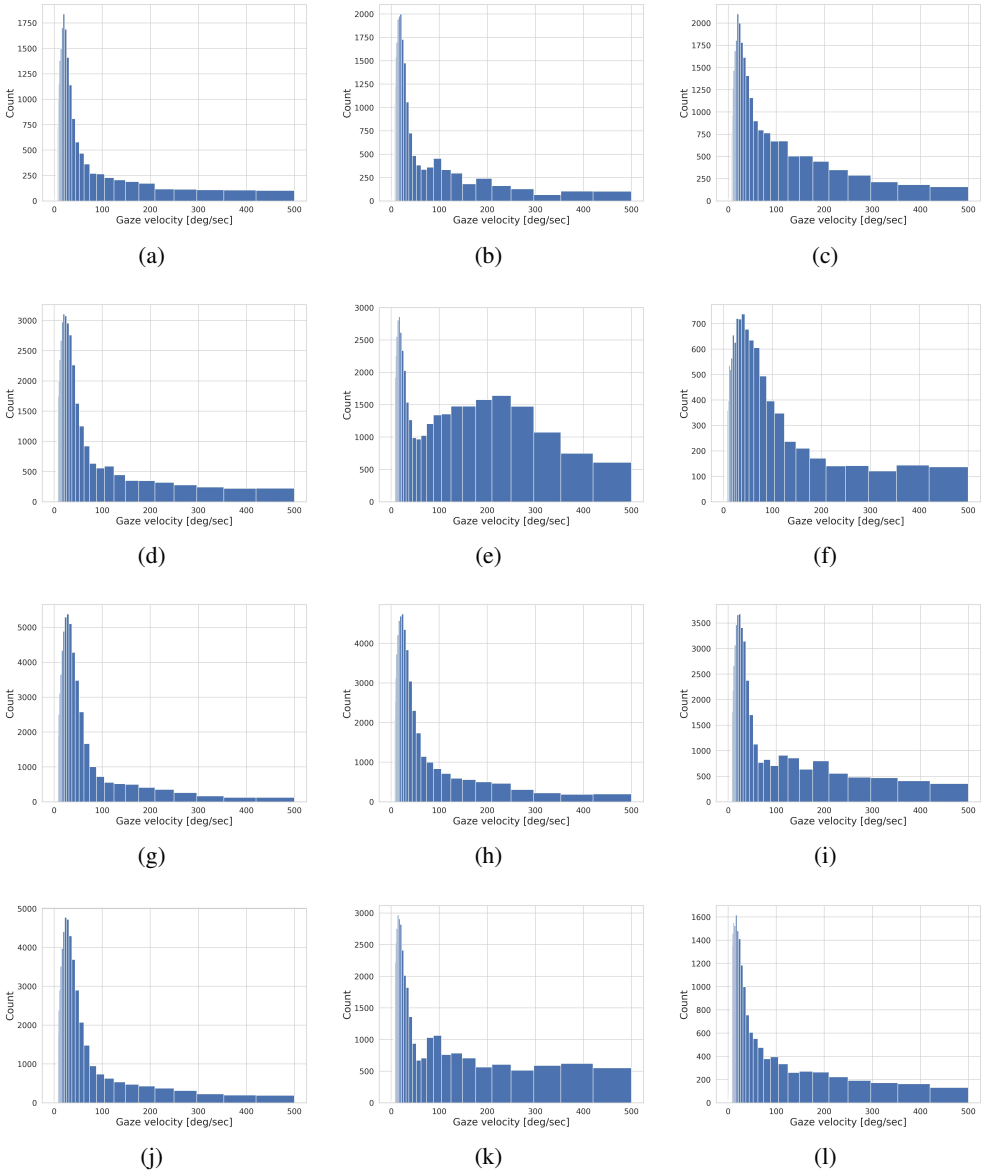


Figure 4: Gaze velocity distribution of randomly selected 12 participants with subject IDs: (a) 1, (b) 3, (c) 5, (d) 12, (e) 13, (f) 14, (g) 20, (h) 25, (i) 29, (j) 33, (k) 39, (l) 40; calculated from Pupil-Core data

Here, the temporal threshold for half-life decay, low-resolution factor and noise threshold are $100ms$, 4 and 1 respectively.

- Refractory period filter (RP): An event is filtered if there is another recent event in the past (in the same pixel) such that the time difference between the present and past events are less than a set-threshold. This filter attempts to neglect the burst of events in the same pixel location. Here, the refractory period is set to be $100ms$.

Through the results in Table 1, it could be perceived that more events are discarded in the interchanging lighting settings (which includes the poor lighting as well) using both LTN and FD filters which validates our earlier premise: poor lighting adds more measurement noise and spatial disparities. Further, the result from RP filter suggests that the collected event data recordings from near-eye movements consist of event bursts in general and thus, no significant impact by that filter on both lighting conditions.

Table 1: Noise analysis between default lighting and interchanging lighting in conventional lab settings. Here, the mean reduction factor (i.e., $\frac{\text{number of discarded events}}{\text{number of all events}}$) from each filtering method is considered.

	Filter		
Type of settings	LTN	FD	RP
Lab setting with default lighting	0.80	0.68	0.95
Interchanging lighting	0.86	0.73	0.95

We further analyse the event data qualitatively as illustrated in Figure 8 where we observe blurred edges in eye structures during poor lighting conditions compared to default lighting (i.e, 65 Lux) conditions which highlights the need for accounting different lighting conditions in event-based eye tracking as a realistic challenge for a dynamic dataset. In addition, we observe a rapid noise accumulation in the event frames in the light transition periods which shows the impact of environmental noise for event-based eye tracking in a realistic setting.

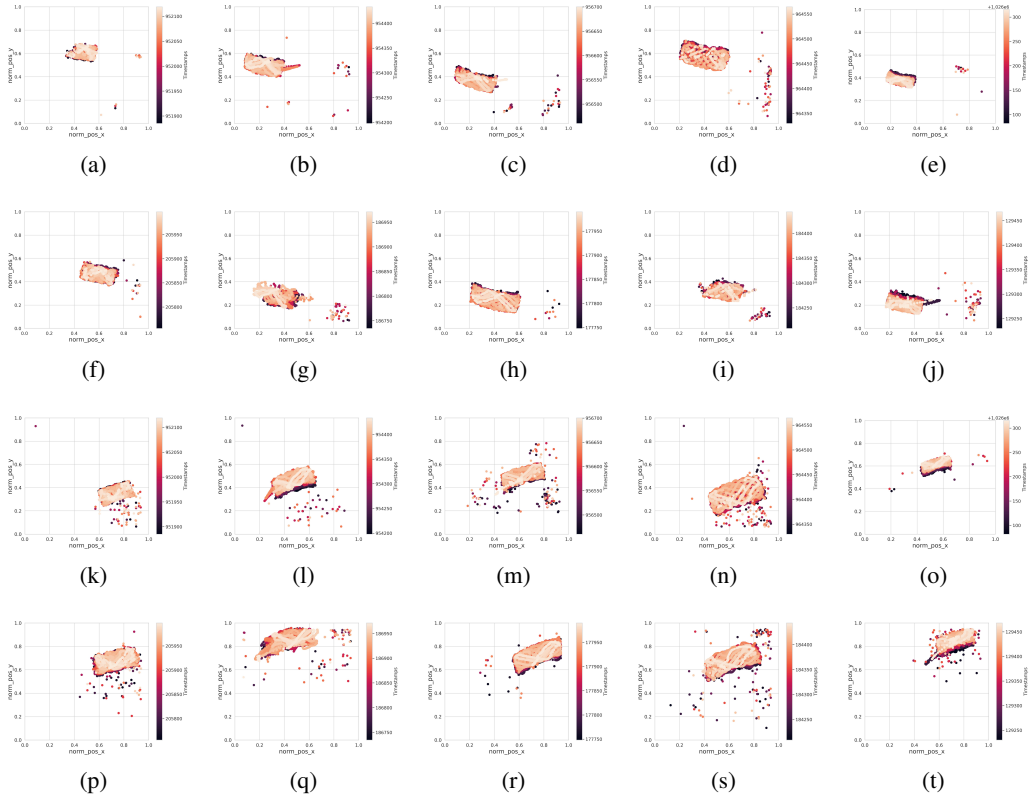


Figure 5: Effect of ambient lighting: spatial pupil coordinates distribution of randomly selected 5 participants for subject IDs 5, 6, 7, 10 and 20 respectively. First row - Right eyes under conventional lab setting with default (65 Lux) lighting, Second row - Right eyes under interchanging (between 65 and 8 Lux) lighting, Third row - Left eyes under conventional lab setting with default (65 Lux) lighting, Fourth row - Left eyes under interchanging (between 65 and 8 Lux) lighting; calculated from Pupil-Core data

2.3 Effect of Participant Motion

We further investigate the effect of participant mobility for both event data and Pupil-Core data while focusing on its impact on the eye movements. In terms of Pupil-Core data, as depicted in figures, the spatial scattering of both pupil coordinates (Figure 9) and gazes (Figure 11) are notably higher than the default seating condition, crucially due to (a) the need of participant to be attentive of

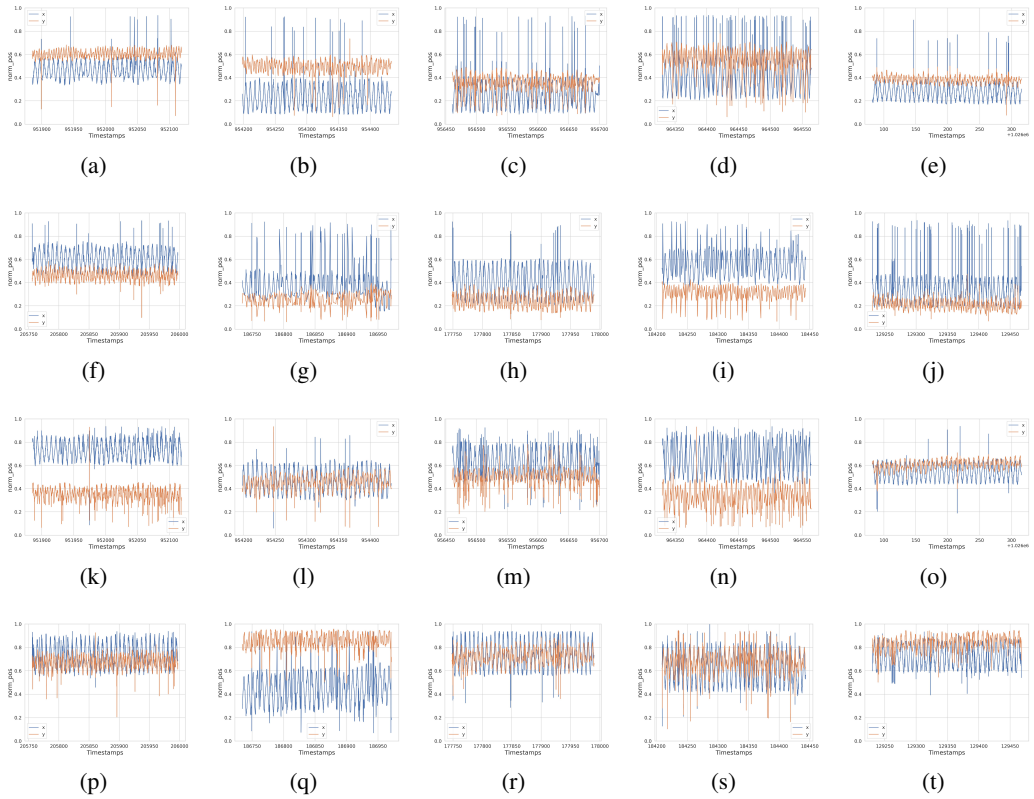


Figure 6: Effect of ambient lighting: temporal pupil coordinates distribution of randomly selected 5 participants for subject IDs 5, 6, 7, 10 and 20 respectively. First row - Right eyes under conventional lab setting with default (65 Lux) lighting, Second row - Right eyes under interchanging (between 65 and 8 Lux) lighting, Third row - Left eyes under conventional lab setting with default (65 Lux) lighting, Fourth row - Left eyes under interchanging (between 65 and 8 Lux) lighting; calculated from Pupil-Core data

both external environment (including walking and surrounding objects) and the visual stimuli when the task is being performed and thus the visual attention is notably divided, and (c) measurement noise is higher in participant mobility due to dynamic conditions. Further, through the Figure 10, it could be perceived that the temporal variation of pupil references, when the participant is moving, is much higher and distorted than when the participant is seated due to the divided visual attention as mentioned above.

With regard to the effect of participant mobility on collected event data, we follow the same noise analysis approach we present in section 2.2 which is for ambient light changes.

Table 2: Noise analysis between default seating settings and user mobility settings. Here, the mean reduction factor (i.e., $\frac{\text{number of discarded events}}{\text{number of all events}}$) from each filtering method is considered.

	Filter		
Type of settings	LTN	FD	RP
Lab setting with user being seated	0.82	0.69	0.94
User is moving freely	0.84	0.73	0.97

As per the results reported in Table 2, more events are flagged as noise across all the filter methods which is intuitive to observe, given that the participant is moving freely. However, through the results from both Tables 1 and 2, it could deduce that the impact of ambient light changes is slightly higher than that of mobility due to (1) the collected event data during the participant mobility is from an

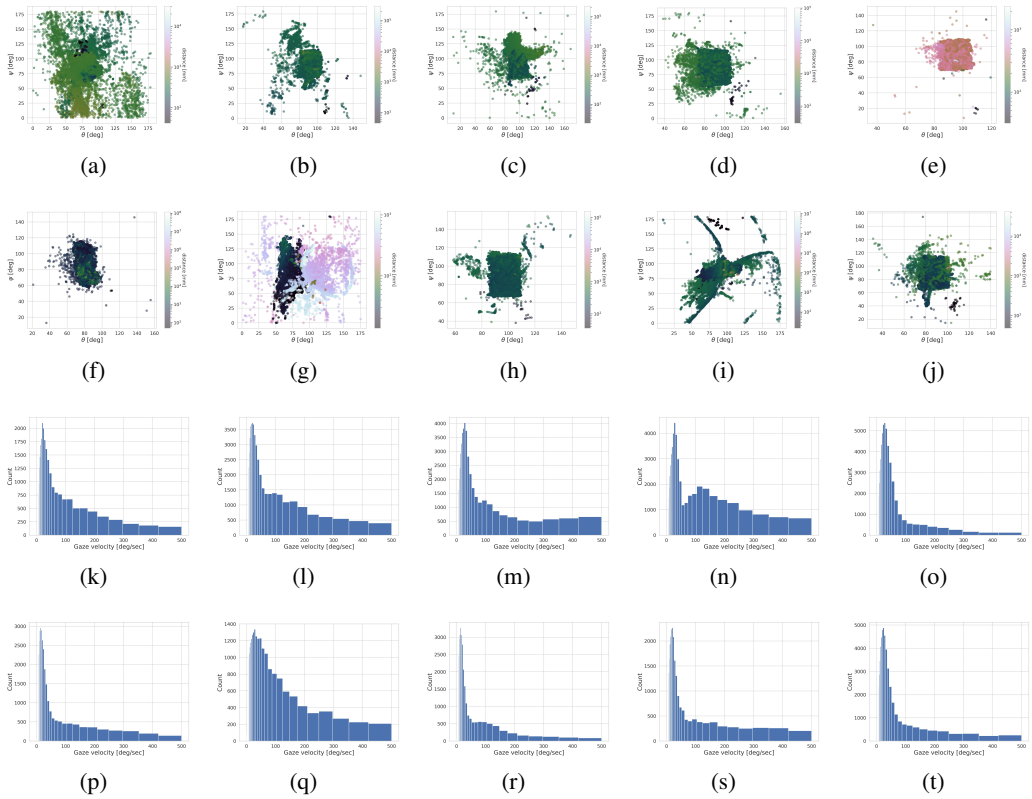


Figure 7: Effect of ambient lighting: gaze angle and velocity distribution of randomly selected 5 participants for subject IDs 5, 6, 7, 10 and 20 respectively. First row - Gaze angles under conventional lab setting with default (65 Lux) lighting, Second row - Gaze angles under interchanging (between 65 and 8 Lux) lighting, Third row - Gaze velocities under conventional lab setting with default (65 Lux) lighting, Fourth row - Gaze velocities under interchanging (between 65 and 8 Lux) lighting; calculated from Pupil-Core data

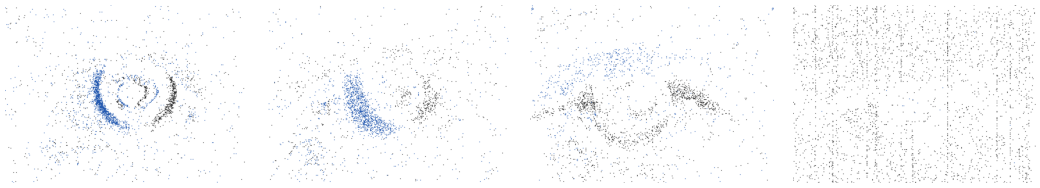


Figure 8: Effect of ambient lighting on event data: Here, the two colors represent positive and negative polarity of events (a) during default lighting condition, (b) and (c) during the poor lighting condition and (d) induced noise when the lighting setting transitions from one setting to the other; extracted from event data

indoor setting and (2) the event camera setup is fixed in position with respect to the eye (using a custom-built head-mounted device which secured around the forehead using a Velcro133 fastener).

Similar to ambient light changes, we qualitatively analyze the effects of participant motion as depicted in Figure 12. Unlike ambient light changes, we do not observe a drastic change in object edges since we experiment in an indoor setting with nearly-constant lighting condition. However, there is a noticeable increment in background noise during the participant mobile setting due to motion-related noise (such as background object noise which are relatively moving with respect to the user, change of direction in illumination on eyes etc.) added to event data.

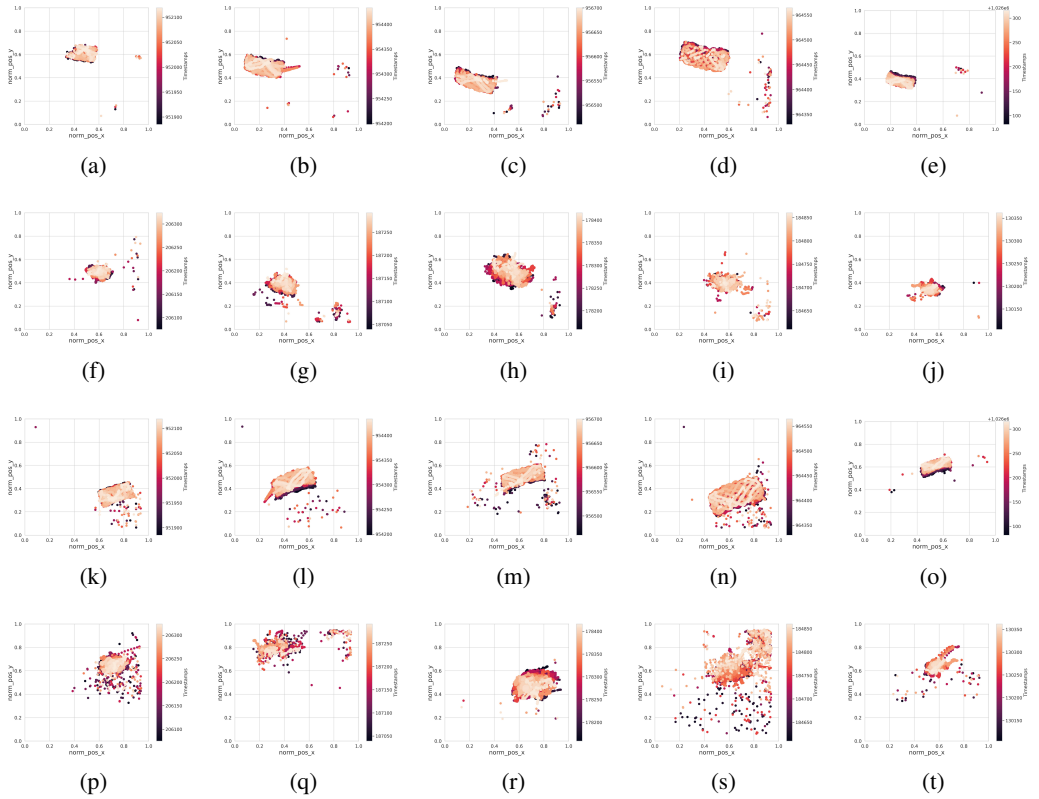


Figure 9: Effect of participant mobility: spatial pupil coordinates distribution of randomly selected 5 participants for subject IDs 5, 6, 7, 10 and 20 respectively. First row - Right eyes under conventional lab setting with participant being seated, Second row - Right eyes under participant is freely moving, Third row - Left eyes under conventional lab setting with participant being seated, Fourth row - Left eyes under participant is freely moving; calculated from Pupil-Core data

2.4 Effect of Head Movements and Eye Rotation Degree

Typically, human gaze movements are influenced by the effect of head movements i.e., the relationship between eye position and gaze direction is not static but rather dynamic, depending on the head's orientation relative to the stimuli [11]. This relationship has not proven to be material in existing datasets [1, 41, 35] since these were collected using fixed-head setups while restricting the natural head movements. Further, to the best of our knowledge, it is also not physiologically feasible to track stimuli using only head movements due to the vestibulo-ocular reflex. This reflex involuntarily stabilizes the human visual field and retinal image during head movements by inducing compensatory eye movements in the opposite direction, making it impractical to isolate head movement alone for tracking purposes [10]. Therefore, even under unrestricted head movement settings as in *EyeGraph* dataset collection, eye movements remain the primary mechanism for visual exploration. Head movements serve a complementary role, extending or adjusting the visual range, but they naturally coordinate with eye movements, often following or aligning with them to track stimuli effectively. This tandem coordination ensures efficient and seamless visual tracking [11, 19].

We observe these coordinated eye and head movements in our dataset as well, specifically through the dense pupil coordinates, gaze angle, and velocity distribution with varying fields of view given that the participants' head movements are unrestricted and natural. For researchers interested in incorporating head movement data into their analysis, our *EyeGraph* event data provides both inertial sensor measurements and external trigger data from the event camera. In addition, our *EyeGraph* Pupil-Core data has been collected by accounting head movements through a robust 3D eye-model-based calibration process. This process involved (1) adjusting the positions and angles of the three embedded cameras on the sliding arms of the Pupil-Core eye tracker to ensure optimal views, and (2)

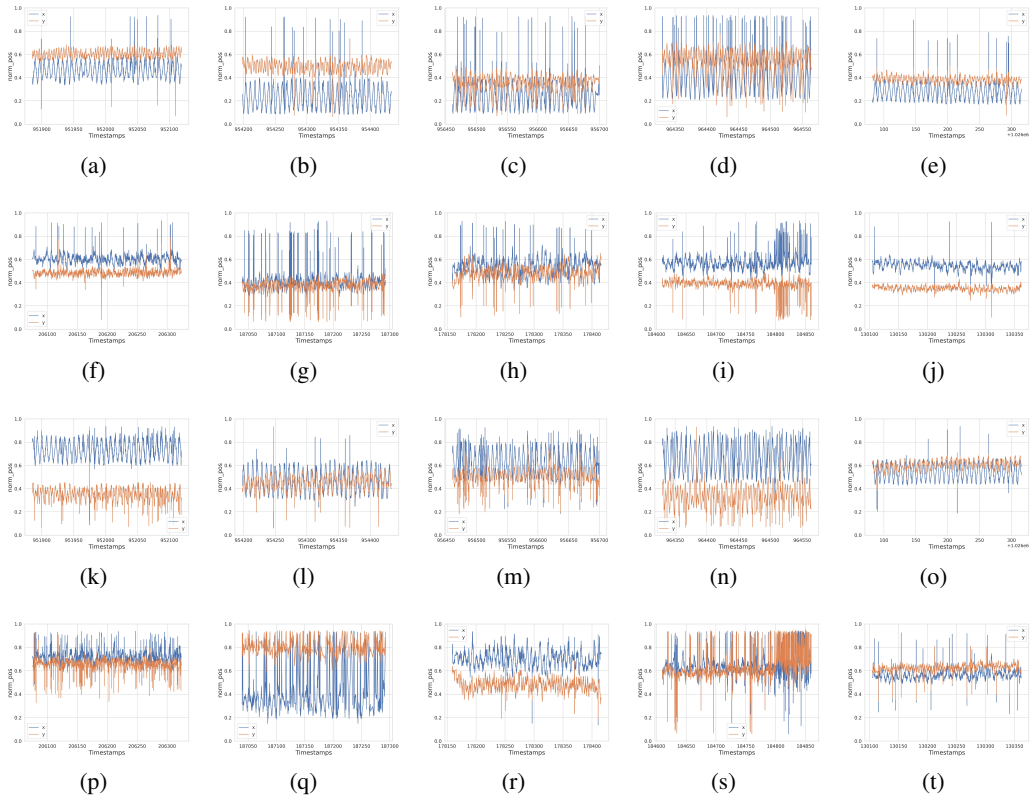


Figure 10: Effect of participant mobility: temporal pupil coordinates distribution of randomly selected 5 participants for subject IDs 5, 6, 7, 10 and 20 respectively. First row - Right eyes under conventional lab setting with participant being seated, Second row - Right eyes under participant is freely moving, Third row - Left eyes under conventional lab setting with participant being seated, Fourth row - Left eyes under participant is freely moving; calculated from Pupil-Core data

calibrating and validating the Pupil-Core eye tracker using a 5-point calibration paradigm, ensuring that the average calibration error remained below $1.5°$, as measured by the Pupil Labs software. In Figure 13, we include the empirical illustrations of 3D pupil rotation angles and inertial measurements (of a randomly selected sample with the head movement) to further extend our analysis in this regard.

2.5 *EyeGraph* Temporal Resolution

Event cameras record asynchronous events with a sub-microsecond latency. It is noteworthy to highlight that, unlike other frame-based event representations used in previous works [41, 36], our (dynamic) graph-based event representation fully preserves the entire temporal resolution delivered by the event camera and propagates through the rest of the pipeline. Further, the rest of our method including *EyeGraph* graph clustering and tube-based RANSAC model does not overlook or reduce the temporal resolution preserved by event graphs but rather propagates that resolution until the final downstream task: pupil coordinates estimation. Therefore, it is possible for *EyeGraph* to continuously predict the pupil coordinates with the fullest temporal resolution delivered by the event camera stream and is only restricted by the hardware limitations of the utilized event camera i.e., DAVIS346 [15] in which the achievable temporal resolution is $1\mu s$. To further extend our analysis, we plot the distribution of (temporally) consecutive events in event data of a randomly selected subject in Figure 14 to show the actual temporal resolution propagated through our pipeline.

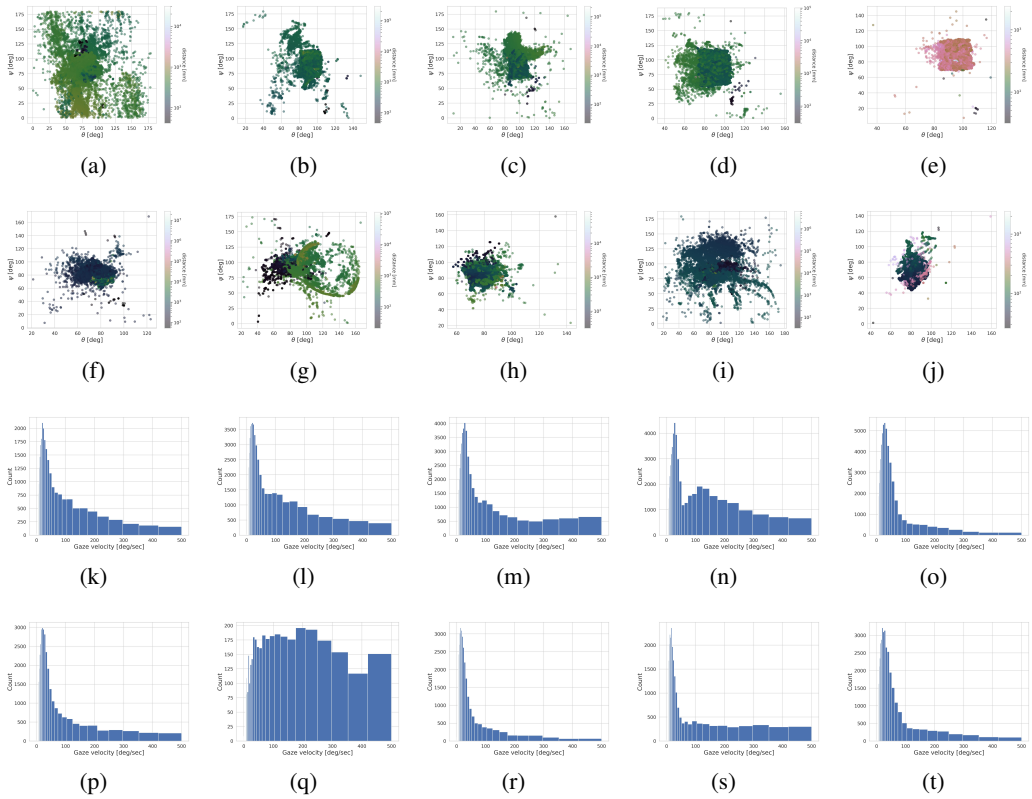


Figure 11: Effect of participant mobility: gaze angle and velocity distribution of randomly selected 5 participants for subject IDs 5, 6, 7, 10 and 20 respectively. First row - Gaze angles under conventional lab setting with participant being seated, Second row - Gaze angles under participant is freely moving, Third row - Gaze velocities under conventional lab setting with participant being seated, Fourth row - Gaze velocities under participant is freely moving; calculated from Pupil-Core data

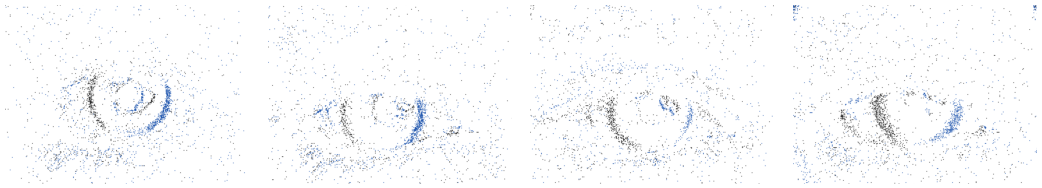
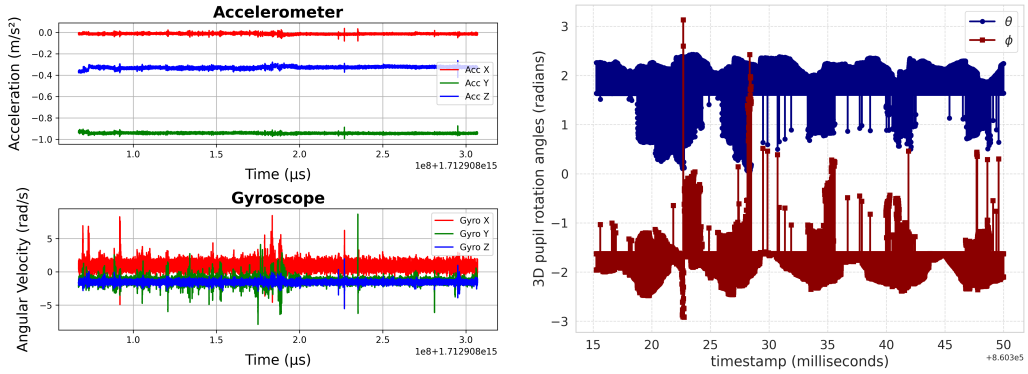


Figure 12: Effect of participant mobility on event data: Here, the two colors represent positive and negative polarity of events (a) during default seating condition, (b), (c) and (d) during the participant is freely moving; extracted from event data

2.6 Human Ground Truth Labelling

Our benchmark method is an *event-only unsupervised* eye tracking approach which does not need to depend on either (i) concurrent RGB frames as existing datasets [1, 41] or (ii) ground truth labels utilized in supervised approaches [35, 36, 4]. However, as reported in the paper (Section 7.4), we provide human-labelled ground truth pupil coordinates for event data collected during the following experimental setups: (1) changing ambient illuminance and (2) user mobility. In our ground truth generation process, we follow an event volume-based annotation procedure for labeling: both Python-based annotation code and the human-labeled ground truths are accessible through the project page. We utilize these human-labelled ground truths to compare the performance of our method using our dataset against the existing supervised methods and clustering approaches. Further, for researchers



(a) (b)

Figure 13: Exemplar illustrations for (a) temporal variations of accelerometer and gyroscope in inertial sensor measurement which are indicative of collected head movements in event data and (b) temporal variations of pupil rotation angles to indicate effective eye movements in unconstrained head motion settings

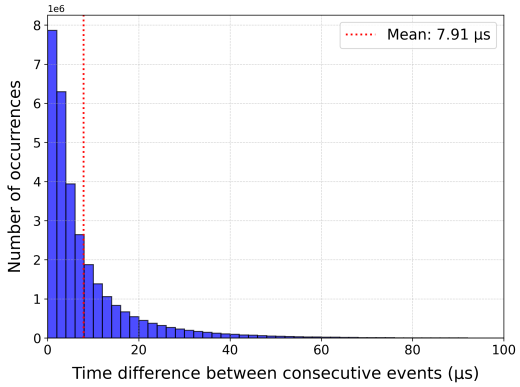


Figure 14: Exemplar illustration for the distribution of (temporally) consecutive events to show the actual temporal resolution propagated through our pipeline. The average time between two consecutive events is approximately 8 micro-seconds with a nominal event frequency of 125,000 Hz in the representative example.

who are interested in utilizing only Pupil-Core data for RGB-based gaze estimation or a related task, our *EyeGraph* dataset provides a rich collection of data (including raw near-eye videos, point of gaze and pupil coordinates, pupil/iris segmentation maps, annotations for fixations/blinks, world-view lookup etc.) – prior to the data collection, the positions and angles of the three embedded cameras on the sliding arms of the tracker are mechanically adjusted. The tracker is calibrated using a 3D eye model-based calibration paradigm to mitigate the effects of potential head movement drifts.

2.7 Potential Negative Impact

We recognize that eye-tracking based techniques, like *EyeGraph* method, may generate (a) inclusivity concerns, especially for individuals with sight impairments, and (b) privacy concerns, especially if emotions and cognitive stresses can be continually captured in workplaces or individuals can be identified only using event eye data. To address the privacy concern(s), we added the following terms and conditions to our custom license for which the users must agree to, before accessing or using *EyeGraph* data. Here, the user/data requestor is referred to as the LICENSEE.

- The LICENSEE will not attempt to identify any individual or institution referenced in *EyeGraph* data.

- The LICENSEE will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in *EyeGraph* data in any publication or other communication.
- If the LICENSEE finds information within *EyeGraph* data that he or she believes might permit identification of any individual or institution, the LICENSEE will report the location of this information promptly by email to the corresponding author, citing the location of the specific information in question.
- Sharing is governed by Creative Commons CC-BY-NC 4.0, but all such redistribution(s) must accompany both licenses.

2.8 Ongoing Efforts for Extending *EyeGraph* Dataset

It is to be noted that our reported experimental setup does not capture all the possible in-the-wild settings and a resulting time-series of variation in illuminance. Instead, it provides a discrete, but representative, set of combinations such as office setting (i.e., seated under standard illuminance), and dimmed illuminance (achieved by intermittently flickering the light switches). We are currently working to collect new instances under varying ambient conditions to update and release new iterations of the dataset. More specifically, we consider the continuous variation in the illuminance scenario and conduct new experiments: the participant is now allowed to roam around freely (while carrying the laptop that shows the same visual stimuli we used earlier) in both indoor and outdoor settings. Further, the walking trail includes physical locations that affect the illuminance, such as shaded trees, moving between two rooms with different lighting conditions, etc. Further, we are collecting both the eye movement data, and close-to-eye illuminance readings using an additional LUX sensor, while an individual subject walks about both indoors and outdoors, over a range of conditions (e.g., early morning, afternoon, night). Our design choice of tracking only a single eye, which ensures that the individual's FoV remains completely unobstructed in one eye, becomes especially salient to prevent risks when exhibiting such natural movements in uncontrolled environments.

3 Supplements for Dynamic Graph Construction

3.1 Event Accumulation

Given that the pixels in event cameras independently trigger events whenever they perceive an intensity change that is above a set threshold, the purpose of event accumulation is to collect a certain number of events such that the collected events are capable of presenting informative scene dynamics. To this end, two major approaches for collecting the events are utilized in the literature: a fixed time interval [18, 40] and a fixed number of events [1, 20]. Even though the utilization of a fixed time interval seems to be trivial, that selection leads to poor performance in downstream tasks due to the scene instability between constant time intervals, i.e. here, eye motion may lead to fewer number of events (if the eye barely moves or does not exhibit any movement) or a higher number of events (as an example, during saccadic eye movements). In contrast, having a fixed number of events ensures a consistent and stable amount of motion given that events are triggered by motion. Further, it also leads to preserve the asynchronous nature of events by allowing to collect events asynchronously. Therefore, in this work, we utilize the latter approach as expressed in Equation 1.

$$E^v = \{e_i^v\}_{i=1}^{C} = \{e_i^v \,|\, t_{e_i^v} \leq t_{e_{i+1}^v}; i \in \{1, C\}; i, C \in \mathbb{N}\} \tag{1}$$

where $C = \{1500, 2000, 4000\}$ is the threshold for the number of events which is heuristically determined such that motion blur or low motion information occurs at minimum.

3.2 Gaussian Mixture Model

Inspired through our empirical observations (as depicted in Figure 15), we utilize a Gaussian mixture model (GMM) on pairwise distances of spatio-temporal event cloud to find the dynamic threshold for radius-graph construction. The utilized GMM with k univariate Gaussian components follows Equation 2.

$$\mathcal{N}(\mu_a, \sigma_a) = \sum_{a=1}^{k} \pi_a \mathcal{N}(x|\mu_a, \sigma_a);\ 1 \leq k \leq c \tag{2}$$

where $0 \leq \pi_i \leq 1$ and $\sum_{a=1}^{k} \pi_a = 1$. In this context, we iteratively run the GMM fitting function for k values in $\{1, c\}$ where $c = 5$ is the maximum number of Gaussian components to be considered, with the objective of reaching the Bayesian information criterion (BIC) [31] to be minimum: $BIC[f(.)] \leq \delta$ where BIC is defined in Equation 2.

$$BIC = \log(N)d + N\log(2\pi) + N\log(\sigma^2) + \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sigma^2} \tag{3}$$

where N, $d = 3k$, σ^2, y_i and $\hat{y}_i$ are the number of samples, the degrees of freedom, the estimate for noise variance, true target and predicted target respectively.

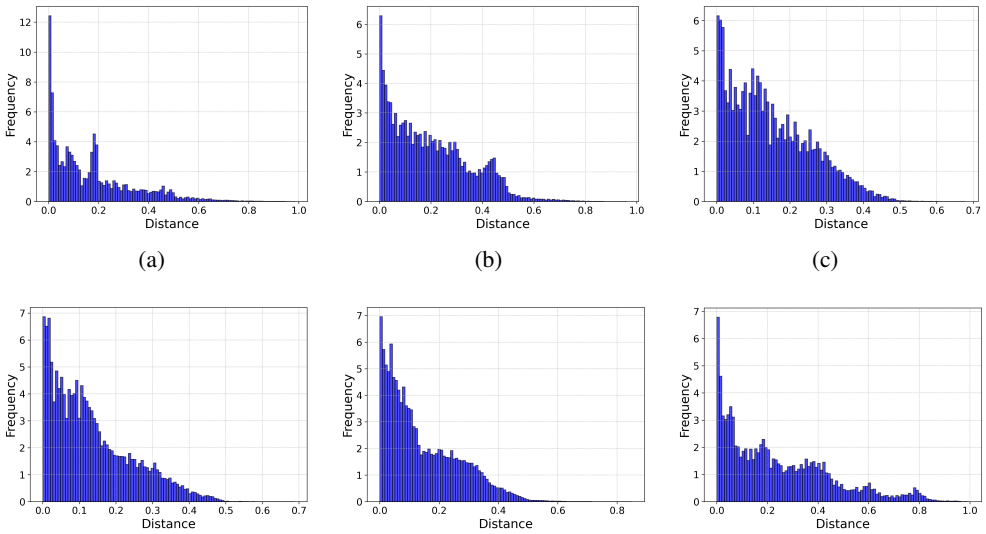


Figure 15: Spatio-temporal distance distributions of 6 randomly selected event volumes

4 Supplements for Topological Graph Clustering

4.1 Overview and Equations

The variational graph autoencoder architecture which we implement in this work follows the Equations 4: message passing rule, 5: encoder, 6: decoder, 7: joint weighted learning objective and 8: Gaussian prior.

$$\mathbf{X}^{(l+1)} = \eta(\widetilde{\mathbf{A}}\mathbf{X}^{(l)}\mathbf{W}^{(l)}) \text{ for } l \in \{0, .., L-1\} \tag{4}$$

where $\widetilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}$ and $\mathbf{D}$ is the diagonal node degree matrix.

$$q(\mathbf{Z} \,|\, \mathbf{X}, \mathbf{A}) = \prod_{i=1}^{n} q(z_i \,|\, \mathbf{X}, \mathbf{A}) = \prod_{i=1}^{n} \mathcal{N}(z_i \,|\, \mu_i, diag(\sigma_i^2)) \tag{5}$$

$$p(\mathbf{A} \,|\, \mathbf{Z}) = \prod_{i=1}^{n}\prod_{j=1}^{n} p(A_{ij} \,|\, z_i, z_j) \text{ with } p(A_{ij} = 1 \,|\, z_i, z_j) = \eta(z_i^\top z_j) \tag{6}$$

$$L = \gamma_1 \mathbb{E}_{q(\mathbf{Z} \,|\, \mathbf{X}, \mathbf{A})}[\log p(\mathbf{A} \,|\, \mathbf{Z})] - \gamma_2 \frac{Tr(\mathbf{B}\mathbf{X}\mathbf{X}^\top)}{2m} - \gamma_3 KL[q(\mathbf{Z} \,|\, \mathbf{X}, \mathbf{A}) \| p(\mathbf{Z})] \tag{7}$$

$$p(\mathbf{Z}) = \prod_{i} \mathcal{N}(z_i \,|\, 0, \mathbf{I}) \tag{8}$$

4.2 Model Hyperparameters

In the dataset preparation, we use a batch size of 32 and to facilitate edge reconstruction-based topological learning process, we split the edges into train, validation and test portions: 85%, 5% and 10% respectively. In terms of model hyperparameters, we utilize RELU activation between layers of the variational graph autoencoder while in learning process, we utilize ADAM optimizer with a constant learning rate of 0.001. Since we utilize balanced iterative reducing and clustering using hierarchies (BIRCH) method [39] for clustering the encoded representations of graphs, which is a tree data structure with the cluster centroids being read off the leaf, we do not need to specify the number of clusters at the front. However, we set the maximum clusters to be 5 following our premise (i.e., distinct anatomical or functional regions of the eye is presented in event data) and empirical evaluations.

5 Supplements for Experiments and Results

5.1 Baseline Methods

The following implementations are followed to generate the results for baseline methods for the datasets: EBV-Eye, 3ET+ and *EyeGraph* .

- Fixed-radius graphs, k-nearest neighbours, nearest neighbour and Furthest point sampling [27] graphs: `https://pytorch-geometric.readthedocs.io/en/latest/_modules/torch_geometric/nn/pool.html`, Accessed: October 31, 2024.
- Gabriel graphs [5]: `https://github.com/GuignardLab/GabrielGraph/`, Accessed: October 31, 2024.
- k-means [2, 21], affinity propagation [12], meanshift [6], spectral clustering [32, 7], DBSCAN [30]: `https://scikit-learn.org/stable/modules/clustering.html`, Accessed: October 31, 2024.
- SBM [26]: `https://github.com/funket/pysbm`, Accessed: October 31, 2024.
- Vanilla graph autoencoder [17]: `https://github.com/pyg-team/pytorch_geometric/blob/master/examples/autoencoder.py`, Accessed: October 31, 2024.
- GSCEventMod [24]: `https://github.com/mondalanindya/ICCVW2021_GSCEventMOD`, Accessed: October 31, 2024.
- DMoN [33]: `https://github.com/pyg-team/pytorch_geometric/blob/master/examples/proteins_dmon_pool.py`, Accessed: October 31, 2024.
- DGI [34]: `https://github.com/pyg-team/pytorch_geometric/blob/master/examples/infomax_transductive.py`, Accessed: October 31, 2024.

5.2 Evaluation Metrics

The detailed explanations on the utilized evaluation metrics for both clustering: Silhouette coefficient, Davis-Bouldin score, Modularity, Conductance; and pupil coordinates estimation: p-accuracy, Euclidean distance, Manhattan distance are presented below. It is to be noted that we use the same notations introduced in the corresponding main paper in the following definitions and only if the notation is not used in the main paper, we introduce them here.

Silhouette coefficient Silhouette coefficient (SC) [28] measures how similar a node is to its own cluster (i.e. cohesion) compared to other clusters (i.e. separation) [37]. In this context, cohesion is measured as the mean Euclidean distance between the node v_i and all other nodes in the same cluster (C_I) (Equation 9 for node v_i) while the separation is calculated as the mean Euclidean distance between the node and all other nodes in another cluster ($C_J; C_J \neq C_I$) as in Equation 10 for node v_i.

$$a(v_i) = \frac{1}{|C_I| - 1} \sum_{v_j \in C_I; v_j \neq v_i} \|v_i - v_j\| \tag{9}$$

$$b(v_i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{v_j \in C_J} \|v_i - v_j\| \tag{10}$$

Following the Equations 9 and 10, SC for one node v_j can be defined as in Equation 11.

$$sc(v_i) = \begin{cases} \frac{b(v_i) - a(v_i)}{\max\{a(v_i), b(v_i)\}} & \text{if } |C_I| > 1 \\ 0 & |C_I| = 1 \end{cases} \tag{11}$$

Using the Equation 11, SC is defined as in Equation 12 [23].

$$SC = \max_c \widetilde{sc}(c) \tag{12}$$

where $\widetilde{sc}(c)$ represents the mean $sc(v_i)$ over all data of the entire dataset for a specific number of clusters c [37]. SC lies between -1 and 1 while higher values are preferred.

Davies-Bouldin score The Davis-Bouldin score (DB) [8] measures the average similarity of one cluster with its most similar cluster where the similarity is calculated through the ratio of within-cluster distances and between-cluster distances.

$$DB = \frac{1}{c} \sum_{I,J=1; I \neq J}^{c} \max \left\{ \frac{\Delta_{itr}(C_I) + \Delta_{itr}(C_J)}{\Delta_{ite}(C_I, C_J)} \right\} \tag{13}$$

where $\Delta_{itr}(.)$ calculates the intra-cluster similarity while $\Delta_{ite}(.)$ calculates the inter-cluster similarity.

Modularity Modularity (Mo) [25] is a graph topology-based measure in which the difference between the fraction of edges within clusters and the expected fraction if edges are distributed randomly is quantified [3]. Mo can be expressed in matrix notation as in Equation 14. Therefore, in summary, a higher value for Mo usually refers to a strong cluster structure.

$$Mo = \frac{1}{2m} \sum_{ij} \mathbf{B} \odot \mathbf{M} = \frac{1}{2m} Tr(\mathbf{B}\mathbf{M}^\top) = \frac{1}{2m} Tr(\mathbf{B}\mathbf{M}) \tag{14}$$

where $\mathbf{M}$ is symmetrical and $n \times n$ of which $M_{ij} = \delta(c_i, c_j)$ such that $\delta(c_i, c_j) = 1$ if and only if $c_i = c_j$ and 0 otherwise. Here, c_i is the cluster for which node v_i is assigned. Further, $\mathbf{B}$ is the modularity matrix such that:

$$B_{ij} = A_{ij} - \frac{dg_i \times dg_j}{2m} = A_{ij} - \frac{\sum_{j=1}^{n} A_{ij} \times \sum_{i=1}^{n} A_{ji}}{2m} \tag{15}$$

where dg_i and dg_j are the degrees of nodes v_i and v_j respectively.

Conductance Conductance (Cd) [32, 38] measures the portion of total edges which goes outside the cluster and therefore, a lower value is preferred. In mathematical terms, Cd can be expressed as in Equation 16.

$$Cd = \frac{|\{(v_i, v_j) \in E : v_i \in C, v_j \notin C\}|}{2(|\{(v_i, v_j) \in E : v_i \in C, v_j \in C\}|) + |\{(v_i, v_j) \in E : v_i \in C, v_j \notin C\}|} \tag{16}$$

for a set of nodes C in V.

p-accuracy This metric is used in the recent works [35] to evaluate the performance of pupil coordinate estimation methods when ground truth is present. It is defined as if the Euclidean distance between the predicted coordinates ($pred_i$) and true coordinates ($true_i$) is within a specified pixel threshold (Th), the prediction is said to be correct and vice versa.

$$p\{Th\} = \frac{1}{N}\sum_{i=1}^{N} f(true_i, pred_i, Th) \text{ with } f(true_i, pred_i, Th) = \begin{cases} 1 & \text{if } \|true_i - pred_i\| \leq Th \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

In this work, we set three pixel thresholds: 10, 5 and 1.

Mean Euclidean and Manhattan distances We utilize two regression metrics as well: Euclidean distance (l_2) and Manhattan distance (l_1), to further evaluate the pupil coordinates estimation.

$$l_2 = \frac{1}{N}\sum_{i=1}^{N} \|true_i - pred_i\|_2 \tag{18}$$

$$l_1 = \frac{1}{N}\sum_{i=1}^{N} |\, true_i - pred_i \,| \tag{19}$$

5.3 Supervised Implementation

For the results reported in Table 4 in the corresponding paper, we annotate the specific event data instances, for pupil coordinates, which are collected during in-the-wild experimental setups: (1) changing ambient illuminance and (2) user mobility. Further, for k-means and DMoN [33] based supervised implementations (marked in Table 4 in the corresponding paper), we design the following pipelines.

- k-means-1: After k-means clustering, we select the cluster with highest density in temporal direction using the below equation, inspired by the following empirical observation: iris and pupil collectively have more density in temporal direction while having low spatial distribution in XY plane. Then, the proposed RANSAC-based pipeline is utilized for pupil coordinates estimation.

$$\begin{aligned} C_{dense(t)\uparrow} &= \frac{\text{Number of nodes in } C_I \,\forall\, I \in \{1, c\}}{\text{Temporal Range}} \\ &= \frac{|C_I| \,\forall\, I \in \{1, c\}}{\max(t_{v_i}) - \min(t_{v_j}) \,\forall\, i, j \in \{1, |C_I|\}} \end{aligned}$$

- k-means-2: Here, we select two clusters with highest densities (from k-means), which are calculated from the above equation, and merge them into one cluster. This is also inspired from our empirical evaluations where we observe the inherent limitation of k-means: in general, the tube structure formed by iris/pupil is divided into two clusters in temporal direction by k-means. Then, we apply the proposed RANSAC-based pipeline for pupil coordinates estimation.
- DMoN: As we report in Table 3 in the paper, DMoN is the most competitive clustering baseline; thus, we also evaluate the system performance only replacing our clustering method by DMoN clustering method within our pipeline i.e, the rest of the pipeline including dynamic graph construction, cluster selection, and RANSAC-based coordinates estimation remained unchanged, exactly following k-means-1 implementation.

6 Datasheet for Dataset

We adopt the recommended documentation framework Datasheets for Dataset [13] for dataset documentation.

6.1 Motivation

1. **For what purpose was the dataset created?**
   The dataset was collected to enable accurate and continuous tracking of eye movements with high temporal resolution in-the-wild.

2. **Who created the dataset?**
   The dataset was created by the authors of this paper from School of Computing and Information Systems, Singapore Management University.
3. **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number?**
   The curation of the dataset is funded by Ministry of Education, Singapore AcRF Tier 1 funding (Grant No: 22-SIS-SMU-044)
4. **Any other comments?**
   No

6.2 Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings;people and interactions between them; nodes and edges)?**
   The data captured by DVS346 are from the right eye of the 40 participants. The data include: event-based data, gray-scale near-eye image frames recorded at 30Hz, inertial sensor measurements, and external trigger data. The corresponding groundtruth was captured by the Pupil Core eye tracker from both the eyes: raw near-eye videos at $\approx$ 100Hz, point of gaze estimations at $\approx$ 100Hz, pupil/iris segmentation, annotations for blinks and fixations.
2. **How many instances are there in total (of each type, if appropriate)?**
   Through 40 data collection instances (where each instance contains raw event and Pupil-Core data as described above per participant), our dataset contains nearly 3.3 billion events and 2 million near-eye grayscale images from DAVIS346 camera, making our dataset the largest of its kind in terms of number of data samples. Further, we have nearly 5 million near-eye video frames from Pupil-Core data. The overall size of the dataset is $\approx 115GB$.
3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
   It contains all the possible instances of the data collected.
4. **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?**
   The dataset consists of all the "raw" data collected from both DVS346 camera, and the Pupil Core eye tracker. In addition we provide ground truth pupil coordinates for event data (from DVS346) for ambient light changes settings and participant mobility settings.
5. **Is there a label or target associated with each instance? If so, please provide a description**
   Labels are only provided for event data (from DVS346) for ambient light-changing settings and participant mobility settings.
6. **Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.**
   No
7. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
   No
8. **Are there recommended data splits (e.g., training, development/validation, testing)?**
   No
9. **Are there any errors, sources of noise, or redundancies in the dataset?**
   No
10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
    Yes, the dataset is self-contained.
11. **Does the dataset contain data that might be considered confidential?**
    No. The dataset is pseudo-anonymized.

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
    No
13. **Does the dataset relate to people?**
    Yes
14. **Does the dataset identify any subpopulations (e.g., by age, gender)?**
    No
15. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**
    No. We explicitly ask all users/data requestors not to exercise such attempts as well via the custom license.
16. **Does the dataset contain data that might be considered sensitive in any way?**
    No
17. **Any other comments?**
    No

6.3 Collection Process

1. **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?** As described in § 1.2 in this material and § 4 of the corresponding main paper, the data was acquired while the participants wearing the neuromorphic event sensor and the Pupil Core eye tracker under three different conditions: controlled lab environment, a room with changing ambient light, and mobility of participants.
2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?**
    The experiment setup, and the characteristics of the collected dataset are detailed in § 1.2 in this material and § 4 of the corresponding main paper.
3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
    The released dataset contains all the possible instances of the data collected.
4. **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
    The participants are from the pool of undergraduate and graduate students and research staff from our school. Few of the participants are members of general public.
5. **Over what timeframe was the data collected?**
    The dataset was collected from 9 April 2024 till May 31 2024.
6. **Were any ethical review processes conducted?**
    Yes. The study was approved by the Institutional Review Board (IRB) of our institute.
7. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
    The data was collected directly from the participants.
8. **Were the individuals in question notified about the data collection?**
    Yes. An explicit consent was obtained, notifying the participants about the data collection and its release to the public and research community.
9. **Did the individuals in question consent to the collection and use of their data?**
    Yes
10. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
    Yes. The participants can contact the PI of this specific work to exclude their data from being analysed or released to the public. The mechanism to revoke the consent is detailed in the participant consent form that was signed by the participants prior to the experiment study.

11. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted?**
    No
12. **Any other comments?**
    No

6.4 Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
   No
2. **Was the"raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
   The collected "raw" data is saved and released.
3. **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point?**
   N/A
4. **Any other comments?**
   No

6.5 Uses

1. **Has the dataset been used for any tasks already?**
   Yes. There are a few ongoing works that use the dataset. However, the dataset will not be claimed as a contribution in those works.
2. **Is there a repository that links to any or all papers or systems that use the dataset?**
   Not currently.
3. **What (other) tasks could the dataset be used for?**
   The dataset can be used to understand latent features that affect human mental health and cognition. Further, it can be used to enhance the quality of content delivery (e.g., foveated rendering) in mixed-reality environments.
4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
   No
5. **Are there tasks for which the dataset should not be used?**
   Yes – the dataset can not be used for commercial purposes.
6. **Any other comments?**
   No

6.6 Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
   The dataset is only available for research purposes.
2. **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?**
   The dataset and the corresponding source code for the benchmark will be publicly available through a project webpage: `https://eye-tracking-for-physiological-sensing.github.io/eyegraph/`.
3. **When will the dataset be distributed?**
   The dataset will be distributed at the time of this paper publishing.
4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
   The dataset will be distributed under two licenses: Creative Commons CC-BY-NC 4.0 and our custom license. Both licenses are made publicly available on our project webpage.

5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
   No
6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
   No

6.7 Maintenance

1. **Who is supporting/hosting/maintaining the dataset?**
   The dataset is maintained by the Human-Machine Collaborative Systems Lab of School of Computing and Information Systems, Singapore Management University.
2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
   The corresponding author (second author) of the paper can be contacted via provided email.
3. **Is there an erratum?**
   No
4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
   Yes. New iterations, augmentations or any other updates will be made publicly available via the project webpage: `https://eye-tracking-for-physiological-sensing.github.io/eyegraph/`.
5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
   No
6. **Will older versions of the dataset continue to be supported/hosted/maintained?**
   N/A
7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
   Yes. The proposed benchmark and the dataset can be accessed from the GitHub repository (at the point of publishing). By adequately citing the corresponding main paper, other researchers can build novel techniques that can be validated using our dataset. However, contributing more instances to the dataset by collecting new data by external researchers may not be covered by the approval we obtained from our IRB.
8. **Any other comments?**
   No

References

[1] A. N. Angelopoulos, J. N. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein. Event-based near-eye gaze tracking beyond 10,000 Hz. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2577–2586, 2021.

[2] D. Arthur, S. Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035, 2007.

[3] A. Bhowmick, M. Kosan, Z. Huang, A. Singh, and S. Medya. DGCLUSTER: A neural framework for attributed graph clustering via modularity maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11069–11077, 2024.

[4] Q. Chen, Z. Wang, S.-C. Liu, and C. Gao. 3et: Efficient event-based eye tracking using a change-based convlstm network. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–5. IEEE, 2023.

[5] J. Choo, R. Jiamthapthaksin, C. S. Chen, O. U. Celepcikay, C. Giusti, and C. F. Eick. MOSAIC: A proximity graph approach for agglomerative clustering. In *Data Warehousing and Knowledge Discovery: 9th International Conference, DaWaK 2007, Regensburg Germany, September 3-7, 2007. Proceedings 9*, pages 231–240. Springer, 2007.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[7] A. Damle, V. Minden, and L. Ying. Simple, direct and efficient multi-way spectral clustering. *Information and Inference: A Journal of the IMA*, 8(1):181–203, 2019.

[8] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[9] P. Duan, Z. W. Wang, X. Zhou, Y. Ma, and B. Shi. EventZoom: Learning to denoise and super resolve neuromorphic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12824–12833, 2021.

[10] M. Fetter. Vestibulo-ocular reflex. *Neuro-Ophthalmology*, 40:35–51, 2007.

[11] E. G. Freedman. Coordination of the eyes and head during visual orienting. *Experimental brain research*, 190:369–387, 2008.

[12] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315 (5814):972–976, 2007.

[13] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[14] inivation. dv-processing, 2024. URL `https://dv-processing.inivation.com/`.

[15] Inivation AG. *DAVIS346b*. Available at `https://inivation.com/wp-content/uploads/2021/08/2021-08-iniVation-devices-Specifications.pdf`, Accessed: October 31, 2024.

[16] M. Kassner, W. Patera, and A. Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pages 1151–1160, 2014.

[17] T. N. Kipf and M. Welling. Variational graph auto-encoders, 2016.

[18] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.

[19] M. F. Land. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, 25(3):296–324, 2006.

[20] N. Li, M. Chang, and A. Raychowdhury. E-gaze: Gaze estimation with event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[21] S. Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

[22] M. B. Miled, W. Liu, and Y. Liu. Adaptive unsupervised learning-based 3D spatiotemporal filter for event-driven cameras. *Research*, 7:0330, 2024.

[23] S. Modak. Finding groups in data: an introduction to cluster analysis: authored by leonard kaufman and peter j. rousseeuw, john wiley and sons, 2005, isbn: 0-47-1-73578-7, 2023.

[24] A. Mondal, J. H. Giraldo, T. Bouwmans, A. S. Chowdhury, et al. Moving object detection for event-based vision using graph spectral clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 876–884, 2021.

[25] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[26] T. P. Peixoto. Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1):012804, 2014.

[27] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[28] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[29] S. Schaefer, D. Gehrig, and D. Scaramuzza. AEGNN: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12371–12381, 2022.

[30] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42 (3):1–21, 2017.

[31] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978. doi: 10.1214/aos/1176344136. URL `https://doi.org/10.1214/aos/1176344136`.

[32] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[33] A. Tsitsulin, J. Palowitch, B. Perozzi, and E. Müller. Graph clustering with graph neural networks. *Journal of Machine Learning Research*, 24(127):1–21, 2023.

[34] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm. Deep graph infomax. *ICLR (Poster)*, 2(3):4, 2019.

[35] Z. Wang, C. Gao, Z. Wu, M. V. Conde, R. Timofte, S.-C. Liu, Q. Chen, Z.-j. Zha, W. Zhai, H. Han, et al. Event-based eye tracking. AIS 2024 challenge survey. *arXiv preprint arXiv:2404.11770*, 2024.

[36] Z. Wang, Z. Wan, H. Han, B. Liao, Y. Wu, W. Zhai, Y. Cao, and Z.-j. Zha. Mambapupil: Bidirectional selective recurrent model for event-based eye tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5762–5770, 2024.

[37] Wikipedia contributors. Silhouette (clustering) — Wikipedia, the free encyclopedia, 2024. URL `https://en.wikipedia.org/w/index.php?title=Silhouette_(clustering)&oldid=1217091312`. [Online; accessed 10-June-2024].

[38] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, pages 1–8, 2012.

[39] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.

[40] T. Zhang, Y. Shen, G. Zhao, L. Wang, X. Chen, L. Bai, and Y. Zhou. Swift-Eye: Towards anti-blink pupil tracking for precise and robust high-frequency near-eye movement analysis with event cameras. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[41] G. Zhao, Y. Yang, J. Liu, N. Chen, Y. Shen, H. Wen, and G. Lan. EV-Eye: Rethinking high-frequency eye tracking through the lenses of event cameras. *Advances in Neural Information Processing Systems*, 36, 2024.