

Local Optical Flow for Eye Movement Event Detection in Head-Mounted Setups

Abdulrahman Mohamed Selim*
Interactive Machine Learning (IML)
German Research Center for Artificial
Intelligence (DFKI)
Saarbrücken, Germany
Applied Artificial Intelligence (AAI)
University of Oldenburg
Oldenburg, Germany
abdulrahman.mohamed@dfki.de

Omair Shahzad Bhatti*
Interactive Machine Learning (IML)
German Research Center for Artificial
Intelligence (DFKI)
Saarbrücken, Germany
omair_shahzad.bhatti@dfki.de

Lukas Wilde
Saarland Informatics Campus
Saarland University
Saarbrücken, Germany
s8luwild@stud.uni-saarland.de

Khue Minh Pham
Interactive Machine Learning (IML)
German Research Center for Artificial
Intelligence (DFKI)
Saarbrücken, Germany
Saarland Informatics Campus
Saarland University
Saarbrücken, Germany
minh_khue.pham@dfki.de

Michael Barz
Interactive Machine Learning (IML)
German Research Center for Artificial
Intelligence (DFKI)
Saarbrücken, Germany
michael.barz@dfki.de

Daniel Sonntag
Interactive Machine Learning (IML)
German Research Center for Artificial
Intelligence (DFKI)
Saarbrücken, Germany
Applied Artificial Intelligence (AAI)
University of Oldenburg
Oldenburg, Germany
daniel.sonntag@dfki.de

Abstract

Detecting eye movement events is essential for analysing visual behaviour and enabling gaze-based interaction. However, many algorithms are designed for stationary eye trackers and struggle in dynamic head-mounted settings, particularly for smooth pursuit detection. In this study, we extend a recent multi-step framework for head-mounted eye movement event detection by replacing globally averaged scene motion with gaze-centred local optical flow extracted using the NeuFlow v2 optical flow estimator. We evaluate this idea in two ways: with a threshold-based approach, and with a unified Random Forest classifier that uses local-flow features to capture object-relative motion. Using the Gaze-in-Wild dataset, and compared with established baselines from the literature, the proposed classifier improves smooth pursuit event-level F_1 from 27.2% (the strongest baseline) to 44.9%, and achieves the highest event-level weighted average ($F_1 = 77.7\%$), while remaining competitive for fixation and saccade detection. We also demonstrated real-time feasibility by integrating our method into a signal processing pipeline with ≈ 40 ms end-to-end latency at 30 FPS on a desktop GPU setup. Despite relying on a single dataset, we show that local optical flow improves head-mounted pursuit detection.

*These authors contributed equally to this work.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

ICMI '26, Napoli, Italy

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2318-6/26/10

<https://doi.org/10.1145/3776574.3831139>

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Ubiquitous and mobile devices**; • **Computing methodologies** → Supervised learning.

Keywords

Head-Mounted Eye Tracking, Eye Movement Event Detection, Smooth Pursuit, Local Optical Flow

ACM Reference Format:

Abdulrahman Mohamed Selim, Omair Shahzad Bhatti, Lukas Wilde, Khue Minh Pham, Michael Barz, and Daniel Sonntag. 2026. Local Optical Flow for Eye Movement Event Detection in Head-Mounted Setups. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3776574.3831139>

1 Introduction

Eye tracking records users' gaze positions and eye movements, enabling researchers to quantitatively study visual behaviour and underlying cognitive processes [8, 10, 16]. While early eye-tracking research was largely confined to stationary, screen-based setups, head-mounted eye trackers extend these capabilities to dynamic real-world environments. They support multimodal interaction in virtual and augmented reality (VR/AR), where gaze can complement speech, gesture, or controller input [12, 18, 20]. Across these settings, reliable detection of eye-movement events, most notably fixations, saccades, and smooth pursuits, is a prerequisite for interpreting visual behaviour and driving interactive systems. Fixations correspond to periods in which eyes remain relatively still, saccades are rapid gaze shifts between fixation points, and smooth pursuits

are slower, continuous eye movements that stabilise the image of a moving object on the fovea [10].

Event detection is substantially harder with head-mounted eye trackers than with stationary remote systems [10]. Head and body motion, dynamic scenes, and changing illumination can introduce motion blur, alter apparent scene motion, weaken gaze-object correspondence, and reduce pupil-detection reliability [6, 14]. As a result, gaze velocity alone is no longer a reliable proxy for eye motion relative to the attended object, which limits traditional threshold-based detectors [24]. In addition, head rotation elicits the vestibulo-ocular reflex (VOR), which produces compensatory eye movements that stabilise gaze in world coordinates [1]. Distinguishing these reflexive movements from voluntary smooth pursuit remains a major source of classification error.

Prior work has addressed these challenges through two main approaches: learning-based and rule-based. **Learning-based methods** model spatio-temporal gaze behaviour and can achieve strong performance in dynamic settings [13, 17]. However, they typically require large amounts of annotated data, may depend on additional sensors such as inertial measurement units (IMUs), and often introduce increased computational cost or latency. By contrast, **rule-based methods** remain attractive because they are lightweight, interpretable, and easy to deploy online. In a recent study, Drews et al. [6] proposed a multi-step framework that augments classical Identification by Velocity Threshold (I-VT) and Dispersion Threshold (I-DT) algorithms [21] using optical flow derived from the scene camera, combined with joint parameter tuning. While this approach improves fixation detection under head motion, it relies on globally averaged optical flow and does not address smooth pursuit detection. Nevertheless, Drews et al. [6] suggest that local optical flow around the gaze point may provide additional discriminative information for this event type.

This gap motivates our study. Our central hypothesis is that local optical flow near the gaze point can help distinguish smooth pursuit from fixation and VOR-driven gaze stabilisation. During pursuit, the eyes follow a moving target whose motion may differ from the dominant background motion or head-induced scene motion. Therefore, capturing this distinction requires motion information local to the gaze location rather than global scene averages. To test this hypothesis, we replace the globally averaged optical flow with dense, gaze-centred local optical flow estimated from the scene camera, and evaluate two approaches designed for online operation using only gaze and scene camera streams: 2-THRESHOLDS directly extends Drews et al. [6] by introducing an additional threshold for smooth pursuit detection; and UNIFIED CLASSIFIER extracts gaze and local optical-flow features and uses them in a machine learning (ML) event classifier.

This work is guided by the following research questions:

- RQ1: Does local optical flow around the gaze point improve fixation and saccade detection compared with gaze-only and global-motion baselines?
- RQ2: Can local optical flow improve smooth pursuit detection in head-mounted eye tracking?
- RQ3: What end-to-end latency and throughput do the proposed approaches achieve, and are they sufficient for online real-time use?

2 Related Work

Classic rule-based eye-movement event detectors, such as I-VT and I-DT [21], remain widely used for their simplicity, interpretability, and low computational cost. I-VT identifies events from gaze velocity, whereas I-DT uses spatial dispersion within a temporal window. In controlled, screen-based experiments, these methods provide reliable baseline performance. However, in head-mounted eye tracking, head motion, scene dynamics, and sensor noise make detection less robust and can fragment events. In addition, both methods only distinguish between saccades and non-saccadic gaze events and do not explicitly separate smooth pursuits from fixations. Andersson et al. [2] evaluated ten threshold-based event detectors across static and dynamic stimuli and showed that detection quality depends strongly on stimulus dynamics and threshold choice. In particular, adaptive thresholding was found to stabilise performance in dynamic scenes, where head and background motion introduce additional artefacts.

Building on these insights, Dar et al. [4] proposed REMoDNaV, an extension of velocity-based detection that uses adaptive thresholds and a multi-stage pipeline to classify fixations, saccades, and smooth pursuits. While this approach improves robustness on dynamic video stimuli, it relies on recording-level statistics and temporal chunking, making it primarily suited for offline analysis and is not directly usable as a single-pass online detector. Recently, **Drews et al. [6]** proposed a three-step framework to improve both I-VT and I-DT for head-mounted eye tracking, as follows: (1) compensate head motion by subtracting global scene motion, computed using sparse Lucas-Kanade optical flow [15], from gaze velocity; (2) apply adaptive thresholding that scales with optical flow magnitude to account for the intensity of head movement; and (3) optimise post-processing, e.g., minimum fixation duration and microsaccade-based merging. Although effective for fixation detection under head motion, the framework does not address smooth pursuit detection, relies on globally averaged optical flow, and performs parameter optimisation offline.

In parallel, several studies have argued that threshold-based detectors are not suitable for mobile settings. For example, Startsev et al. [24] noted that simple algorithms, such as I-VT and I-DT, can be flawed when used with mobile eye trackers. Therefore, various event detectors have been proposed that use deep learning to learn discriminative patterns directly from gaze data, rather than relying on handcrafted rules. Early work by Hoppe et al. [11] introduced an end-to-end CNN-based classifier that operates on short, overlapping gaze windows transformed into frequency representations; their model avoids manual threshold tuning and can classify fixations, saccades, and smooth pursuits; however, it relies on sliding windows, which introduce detection latency, and was primarily evaluated on static, remote-tracker data. Other approaches combined convolutional feature extraction with recurrent models to capture longer temporal context. Startsev et al. [23] proposed a 1D CNN-BLSTM architecture that performs well on dynamic screen-based datasets such as GazeCom [5]; however, bidirectional recurrence limits low-latency operation, and the method remains dependent on supervised training and tracker-specific tuning.

Targeting head-mounted eye tracking explicitly, Elmadjian et al. [7] proposed a Temporal Convolutional Network (TCN) that achieved

Table 1: Comparison of representative eye movement event detection approaches. “Partly” indicates limited suitability (e.g., not designed for head-mounted recordings, not evaluated under strict real-time constraints, or known to degrade substantially under head motion). Bold entries indicate desirable properties.

Approach	Online real-time operation	Head-mounted eye tracking suitable	Detects smooth pursuit	No supervised training required	No additional sensors required
I-VT / I-DT [21]	Yes	Partly	No	Yes	Yes
REMoDNaV [4]	No	Partly	Yes	Yes	Yes
1D CNN + BLSTM [23]	No	No	Yes	No	Yes
TCN [7]	Yes	Partly	Yes	No	Yes
GRU (with IMU) [13]	Yes	Yes	Yes	No	No
Steil et al. [25]	Yes	Yes	No	Yes	Yes
ACE-DNV [17]	Partly	Yes	Yes	No	Yes
Drews et al. [6]	Partly	Yes	No	No	Yes
Our method	Yes	Yes	Yes	Partly	Yes

low-latency predictions and strong performance on screen-based benchmarks; however, their head-mounted evaluation focused on relatively static viewing tasks (e.g., watching videos), leaving open questions about performance in more natural, highly dynamic environments. Kothari et al. [13] introduced a GRU-based detector that incorporates head-rotation information from an IMU, reporting improved performance, particularly for smooth pursuit, when head motion is available; while effective, this dependence on additional sensors increases hardware complexity and reduces general applicability. An earlier approach by Steil et al. [25] showed that scene-video cues can also support head-mounted event detection by comparing visual similarity between consecutive gaze-centred scene patches for fixation detection. Building on these two approaches, Nejad et al. [17] proposed ACE-DNV, which combines gaze features, visual similarity, and head/body motion inferred via visual odometry from the scene camera; although this approach supports smooth pursuit detection without an IMU, it relies on multiple computationally intensive components and requires supervised training, and its real-time feasibility was not evaluated under explicit latency constraints.

Table 1 summarises the most representative event detection approaches discussed above that are relevant to our study. Overall, learning-based methods demonstrate that smooth pursuit can be detected in dynamic settings when sufficient temporal or motion information is available. At the same time, they often require supervised training, additional sensors, or increased computational overhead. In contrast, this work investigates whether augmenting lightweight, threshold-based detectors with *local optical flow* around the gaze point can provide sufficient motion context to improve smooth pursuit detection while maintaining real-time performance and avoiding the need for additional sensors.

3 Method

Building on the multi-step framework of Drews et al. [6], we propose two extensions to support smooth pursuit event detection in head-mounted eye tracking: a lightweight rule-based two-threshold

approach (i.e., 2-THRESHOLDS), and a learning-based approach using an ML classifier (i.e., UNIFIED CLASSIFIER). Although the original framework [6] (see section 2) improves fixation detection under head motion using I-VT and I-DT, it does not explicitly address smooth pursuit. Therefore, we propose an adaptation that separates pursuit from fixation-like gaze stabilisation in dynamic scenes while retaining low latency.

Similar to Drews et al. [6], we use optical flow as a proxy for head-induced scene motion. However, instead of sparse flow sampled across the full image and then globally averaged, we compute dense optical flow and extract a gaze-centred local region for analysis. To obtain dense optical flow with low latency, we use NeuFlow v2 [28], an efficient optical flow estimator designed to balance accuracy and throughput. We selected it because it provides dense local motion estimates required for our gaze-centred features while remaining suitable for online processing and operation on edge devices. This addresses a limitation of global averaging, which can suppress motion signals from independently moving objects near the gaze point. These local signals can be informative when the observer fixates or pursues a target whose motion deviates from the dominant background motion. Therefore, for each gaze sample, we extract a gaze-centred patch from the dense flow field and use this local representation for threshold-based detection and to compute features for ML-based classification. Our codebase is publicly available online on [GitHub](#).

3.1 Threshold-based (2-Thresholds)

We follow the same three strategies as Drews et al. [6] with I-VT and I-DT, but introduce an additional threshold on the relative gaze velocity to explicitly detect smooth pursuit. Gaze velocity $\mathbf{v}_{gaze}$ is obtained by differentiating the gaze coordinates; the relative gaze velocity is computed as follows:

$$\mathbf{v}_{rel} = \mathbf{v}_{gaze} - \mathbf{v}_{flow}, \quad (1)$$

where $\mathbf{v}_{flow}$ denotes the local optical flow vector at the gaze point, extracted from the dense NeuFlow v2 flow field. Intuitively, $\mathbf{v}_{rel}$

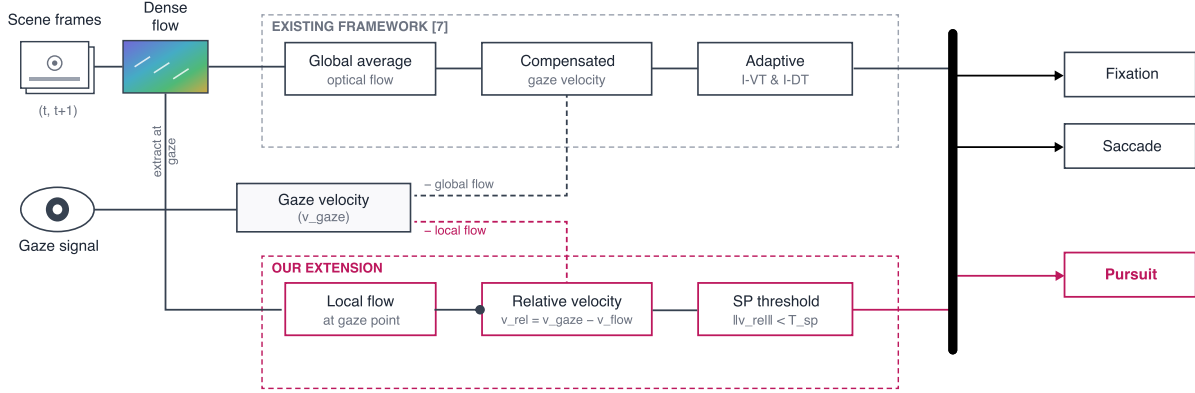


Figure 1: Overview of the 2-THRESHOLDS. The existing framework [6] (top, grey) compensates gaze velocity with globally averaged optical flow and classifies fixations and saccades via adaptive I-VT/I-DT. Our extension (bottom, red) replaces the global average with gaze-centred local optical flow and applies two thresholds on the resulting relative velocity $\|v_{rel}\|$ to additionally detect smooth pursuit.

captures eye motion *relative to the attended scene region*: low $\|v_{rel}\|$ indicates that the eyes move in concert with the local scene, which occurs during both fixations (eyes compensate for head motion via the VOR) and smooth pursuits (eyes track a moving target); high $\|v_{rel}\|$ indicates a saccade, where gaze shifts rapidly relative to the scene. The 2-THRESHOLDS variant partitions $\|v_{rel}\|$ using two thresholds, T_{sp} and T_{sacc} (with $T_{sp} < T_{sacc}$):

- $\|v_{rel}\| > T_{sacc}$: classified as **saccade**,
- $T_{sp} < \|v_{rel}\| \leq T_{sacc}$: classified as **smooth pursuit** candidate,
- $\|v_{rel}\| \leq T_{sp}$: classified as **fixation** (Fi+Fo).

The 2-THRESHOLDS approach assigns sample-level labels and then applies standard event-level post-processing to obtain coherent events. This extension is conceptually simple, remains fully rule-based, and serves as our primary lightweight baseline. Figure 1 provides an overview of our 2-THRESHOLDS approach.

3.2 Unified Classifier

In addition to the 2-THRESHOLDS approach, we introduce the UNIFIED CLASSIFIER approach inspired by ACE-DNV [17]. We use a Random Forest classifier (RF) to predict fixations, saccades, and smooth pursuits. However, rather than relying on visual odometry, patch similarity, or inertial sensors, we derive features directly from gaze signals and local optical flow. Figure 2 shows a simple visualisation for this approach.

We compute *gaze velocity magnitude and direction* from v_{gaze} as our gaze features. Saccades exhibit short, high-velocity peaks, whereas fixations and pursuits yield lower speeds; directional consistency further characterises smooth pursuit. In addition, we compute four optical flow features. First, we reuse the relative gaze velocity magnitude $\|v_{rel}\|$ as a feature, which captures eye motion relative to the attended scene region; we refer to it as *relative motion*. Second, we compute the *foveal-peripheral flow difference*. Let $\mathcal{R}_{fov}$ denote a foveal region of radius r_{in} around the gaze point g ,

and $\mathcal{R}_{per}$ a surrounding peripheral ring with outer radius r_{out} . The mean flow vectors are

$$\bar{v}_{fov} = \frac{1}{|\mathcal{R}_{fov}|} \sum_{p \in \mathcal{R}_{fov}} v_{flow}(p), \quad \bar{v}_{per} = \frac{1}{|\mathcal{R}_{per}|} \sum_{p \in \mathcal{R}_{per}} v_{flow}(p), \quad (2)$$

and the feature is given by

$$\Delta_{flow} = \|\bar{v}_{fov} - \bar{v}_{per}\|_2. \quad (3)$$

This measures whether the attended region moves independently of the background, which could indicate a smooth pursuit.

Third, we compute the *optical flow entropy* within the gaze-centred region. Flow magnitudes and directions are discretised into bins with probabilities p_i , and entropy is defined as

$$H = - \sum_i p_i \log_2 p_i \quad \text{with} \quad \sum_i p_i = 1. \quad (4)$$

Low entropy indicates coherent or near-static motion, whereas high entropy reflects disordered flow patterns typical of saccades or complex scenes.

Finally, we compute the *cosine similarity* between gaze velocity and local optical flow,

$$S_{cos} = \frac{v_{gaze} \cdot v_{flow}}{\|v_{gaze}\|_2 \|v_{flow}\|_2}. \quad (5)$$

During fixations under head motion, gaze velocity aligns strongly with background flow due to the VOR, whereas smooth pursuit and saccades often deviate from this pattern. The classifier produces sample-wise predictions, which are post-processed using standard minimum-duration constraints. By restricting optical-flow computation and feature extraction to gaze-centred regions and avoiding recurrent architectures, the pipeline remains lightweight.

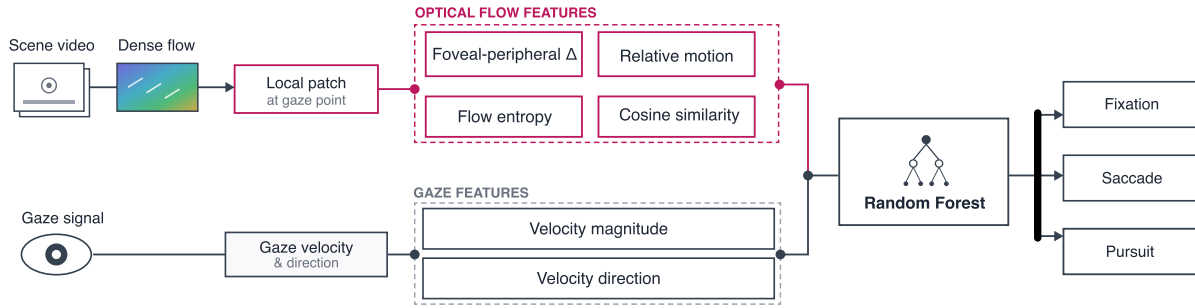


Figure 2: Overview of the UNIFIED CLASSIFIER. Dense optical flow is computed from consecutive scene frames and sampled locally at the gaze point. Gaze and flow features are passed to an RF classifier for sample-wise prediction.

3.3 Implementation Details

We implemented the approach in Python, using the official NeuFlow v2 implementation¹ for dense optical flow. Flow was computed on scene frames resized to 768×432 pixels, rescaled to the original resolution, and then used for feature extraction. We smoothed gaze coordinates with a Savitzky-Golay filter [22], computed point-to-point gaze velocity using each recording’s mean sampling interval, and used a 25 ms window for dispersion-based processing. When multiple gaze samples occurred between consecutive video frames, they were assigned the same flow vector. Additional implementation details and parameter search spaces are provided in the Appendix ??.

Threshold-based methods were tuned by grid search on each fold’s training recordings. The UNIFIED CLASSIFIER used a scikit-learn RF with default settings and random undersampling of over-represented classes within each training fold. For reproducibility, we fixed the NumPy seed and scikit-learn random state to 42, and repeated all experiments with seeds 256 and 999; results are reported in the Appendix ??.

4 Experiments

Since our method directly extends the framework proposed by Drews et al. [6], we used their dataset (denoted DD) to validate our re-implementation and fixation-detection behaviour. We also evaluated pursuit detection on the Gaze-in-Wild (GrW) dataset [13], which was recorded in dynamic everyday environments and explicitly contains smooth pursuit labels. To allow direct comparison with prior published work, we used slightly different experimental designs for each dataset, including cross-validation and event-matching strategies.

The DD dataset contains approximately 14 minutes of annotated head-mounted gaze data recorded with Pupil Invisible glasses [26]. It includes a mixture of relatively static and dynamic tasks, such as supermarket search, driving, playing an instrument, and free exploration. The dataset contains only annotations distinguishing between fixations and gaps, with gaps treated as saccades.

The GrW dataset contains approximately 2 hours and 15 minutes of head-mounted recordings with scene video and gaze annotations across tasks, including navigation, visual search, ball catching,

and tea making. It contains annotations for *fixation*, *following*, *saccade*, *smooth pursuit*, and *blink*. The smooth pursuit annotations occur only in the “Ball Catching” task. Although IMU-based head rotation signals are available, we do not use them to keep the comparison restricted to standard eye-tracker outputs (i.e., gaze + scene video). Following Nejad et al. [17], we select annotations from the most consistent labeller per task and exclude the “Tea Making” task due to its limited number of samples.

4.1 Evaluation metrics

We report sample-level and event-level F_1 scores as the primary performance metrics. Event-level evaluation requires matching predicted events to ground-truth events. For the DD dataset, following Drews et al. [6], we use Intersection-over-Union (IoU) matching with a minimum IoU threshold of 0.7. However, for the GrW dataset, similar to Nejad et al. [17], we apply majority-voting matching, which tolerates short interruptions within longer events. Therefore, sample-level and event-level scores are not directly comparable, because they penalise different types of errors. In particular, majority-voting event matching can still count a fragmented prediction as correct if the dominant label within a ground-truth event is preserved.

4.2 Validation on the Drews and Dierkes Dataset

For DD, we follow the original 5-fold cross-validation protocol reported by Drews et al. [6] at the recording level. In each fold, approximately three static and four dynamic recordings are held out for validation, and the remaining recordings are used for training and parameter tuning. Parameters are selected via grid search on the training recordings by maximising the per-recording averaged event-level fixation F_1 score, and then evaluated on the fold’s held-out recordings.

Following the original work, we treat this dataset as a binary setting, with fixations as the positive class and gaps as the negative class, corresponding to saccades. We compare our results against classical I-VT and I-DT [21] to validate our re-implementation of the multi-step framework, similar to Drews et al. [6]. This dataset is primarily used to verify that our implementation reproduces the fixation-detection behaviour of the reference framework and that swapping global optical flow for our gaze-centred local optical flow does not harm fixation detection.

¹https://github.com/neufieldrobotics/NeuFlow_v2 (accessed 10 April 2026)

Table 2: Sample- (S) and event-level (E) F_1 scores on GrW using leave-one-out cross-validation.

Algorithm	Fi + Fo		Saccade		Pursuit		W. Avg.	
	S	E	S	E	S	E	S	E
TCN	84.2	72.8	48.4	67.2	1.1	0.8	74.6	68.0
ACE-DNV	64.4	66.2	49.2	80.1	12.6	12.7	59.6	71.2
GRU	72.1	74.9	55.8	80.1	28.8	27.2	73.1	74.2
2-THRESHOLDS	87.6	79.6	61.4	80.0	8.0	0.8	78.9	77.1
RF-LOCALFLOW (FULL)	78.2	77.1	52.6	80.5	27.1	44.9	71.8	77.7
RF-LOCALFLOW (-FOVEAL)	77.6	77.1	52.4	80.8	26.8	43.9	71.2	77.8
RF-LOCALFLOW (-ENTROPY)	76.7	76.3	51.6	79.7	25.4	43.6	70.3	76.9
RF-LOCALFLOW (-COS)	78.3	77.1	52.8	80.7	27.8	45.9	71.9	77.8
RF-LOCALFLOW (-RELMOTION)	78.9	77.2	52.8	80.3	27.6	46.8	72.4	77.7
RF (GAZE-ONLY)	55.1	56.4	42.9	76.2	13.4	18.5	51.2	64.8

4.3 Evaluation on Gaze-in-Wild Dataset

For GrW, we followed the same experimental setup as Nejad et al. [17]. We used leave-one-out (LOO) cross-validation at the recording level, motivated by the strong class imbalance and the fact that smooth pursuit labels occur only in the “Ball Catching” task. LOO ensures that pursuit samples remain available in the training split of each fold. We merged *fixation* and *following* into a single class (Fi+Fo), since our lightweight detectors do not explicitly separate these categories, and excluded *blinks*. Here, *following* refers to gaze stabilisation on a stationary world-fixed target during observer motion, whereas *smooth pursuit* refers to tracking a moving target. To address the imbalance between event types, especially smooth pursuit, we applied random undersampling to the training split of each fold. For each over-represented class, we removed a random subset of samples, sized so that class’s count approaches that of the smallest class, rather than discarding whole events, to preserve their temporal structure. Events shorter than three samples were left untouched, and at least one sample per event always survived. This ran independently within the training portion of each leave-one-out fold and never touched the held-out recording.

We compare against three baselines suitable for head-mounted data, GRU [13], TCN [7], and ACE-DNV [17], all re-run under the same random seeds and data splits as our own experiments, with IMU input excluded throughout, to ensure comparability. In each LOO split, the baselines and our UNIFIED CLASSIFIER are trained on the training recordings and evaluated on the held-out recording. The 2-THRESHOLDS method does not use supervised training, but its thresholds and post-processing parameters are selected by grid search on the training recordings of each fold.

4.3.1 2-THRESHOLDS. As our simplest extension of the multi-step framework, the 2-THRESHOLDS introduces a second threshold on the local-flow compensated relative velocity to model smooth pursuit candidates. This experiment tests whether velocity matching alone is sufficient to separate smooth pursuits from fixations and saccades in a head-mounted setting.

4.3.2 Unified classifier and ablation study. We evaluate our RF-based UNIFIED CLASSIFIER to jointly predict Fi+Fo, saccade, and smooth pursuit using gaze features and local optical flow features,

namely relative motion, cosine similarity, foveal/peripheral flow contrast, and flow entropy. To quantify the contribution of local optical flow, we perform an ablation study by progressively removing feature groups. All variants are evaluated with the same LOO protocol and majority-voting event matching.

4.3.3 Real-time feasibility. To answer RQ3, we integrated the UNIFIED CLASSIFIER into the Multisensor Pipeline (MSP) [3] and re-played GrW recordings as a streaming input. We report end-to-end latency from frame arrival to label prediction, and throughput in frames per second (FPS), including a breakdown of optical flow estimation, feature extraction, and classification runtimes.

4.4 Results

4.4.1 Validation on DD. Overall, our results show that the reimplementation of Drews et al. [6] was successful. For the baselines, we obtained sample-level fixation performance of $F_1 = 94$ (I-VT) and $F_1 = 95$ (I-DT), with event-level fixation $F_1 = 54$ (I-VT) and $F_1 = 59$ (I-DT). Adding flow-based motion compensation and post-processing improved event-level performance $F_1 = 77$ (I-VT+F) and $F_1 = 75$ (I-DT+F), consistent with the improvements reported by Drews et al. [6]. Importantly, swapping the original global optical flow with our NeuFlow v2 local optical flow did not degrade fixation detection: event-level fixation F_1 remained unchanged for I-VT+F (77 vs. 77) and changed only slightly for I-DT+F (75 vs. 72.8), while the simpler baselines showed at most small differences (I-VT: 54 vs. 50; I-DT: 59 vs. 59). This indicates that NeuFlow v2 is suitable as a drop-in optical-flow estimator for our gaze-centred motion features while preserving fixation-detection performance on DD.

4.4.2 Evaluation on GrW. Following the same experimental design as Nejad et al. [17], and using only gaze and video data for all methods, Table 2 summarises sample- and event-level F_1 scores on GrW. The 2-THRESHOLDS achieves the highest sample-level F_1 for Fi+Fo (87.6), Saccade (61.4), and the weighted average (78.9), but yields very low Pursuit performance (8.0 sample-level / 0.8 event-level). Among the baselines, GRU achieves the best sample-level Pursuit F_1 (28.8), while TCN remains competitive on Fi+Fo at the sample level (84.2) but performs poorly for Pursuit (1.1 sample-level / 0.8 event-level).

For the UNIFIED CLASSIFIER, the event-level scores of the RF variants are tightly clustered, i.e., RF-LOCALFLOW (FULL) reaches 77.1 (F1+FO), 80.5 (Saccade), and 77.7 (weighted average), each within about one point of the best single-feature variant on that metric. Event-level Pursuit F_1 is 44.9 for the full model and ranges between 43.6 and 46.8 across the single-feature ablations, whereas removing optical-flow features entirely (RF (GAZE-ONLY)) reduces performance across all classes and at both levels, with the weighted average dropping to 51.2 sample-level and 64.8 event-level and pursuit falling to 18.5. In the Appendix ?? has the full ablation results. We evaluated the significance of our results using the two-sided Wilcoxon signed-rank test [27] on the per-fold, event-level F_1 scores, applying the Holm-Bonferroni correction [9] within each comparison family to account for multiple testing; the full table is in the Appendix ???. The analysis confirms that only removing all flow features yields a statistically significant change, while no single-feature removal differs significantly from the full model. A finer event-structure analysis (Appendix ??) shows that the classifier overlaps most pursuit events but tends to fragment them, so the event-level gains would likely benefit from light temporal smoothing, while confirming that the 2-THRESHOLDS baseline fails to produce coherent pursuit events. Overall, these results show that local optical flow contributes little to the already strong threshold-based detection of F1+FO and Saccade, but is important for event-level smooth pursuit detection.

4.4.3 Real-time feasibility. We integrated the RF-LOCALFLOW (FULL) classifier into MSP and streamed GrW recordings at 30 FPS to obtain online, continuous label predictions. On a desktop PC with an AMD Ryzen 9 3900XT (12 cores, 3.8 GHz), an NVIDIA RTX 3080 (10 GB), and 32 GB RAM, the pipeline sustained 30 FPS throughput with ≈ 40 ms end-to-end latency, of which ≈ 30 ms was spent on optical flow, ≈ 3 ms on feature extraction, and ≈ 7 ms on classification. On [GitHub](#), we have a video demonstration of the live pipeline.

5 Discussion

In this study, we extended the framework of Drews et al. [6] by replacing global averaged scene motion with dense local optical flow around the gaze point. We evaluated two ways of using this information: a rule-based 2-THRESHOLDS extension of the original multi-step pipeline, and a single RF-based UNIFIED CLASSIFIER that combines gaze and local-flow features. Overall, the results show the benefits of using richer local motion descriptors, and we discuss these findings with respect to our three research questions.

5.1 RQ1. Local Optical Flow for Fixation and Saccade Detection

To answer RQ1, we compared against two families of baselines: **Gaze-only** detectors that only use gaze signal (i.e., classical I-VT/I-DT on both datasets and the RF (GAZE-ONLY) on GrW), and **Global-motion** detectors that compensate gaze velocity with globally averaged scene flow (i.e., the original I-VT+F / I-DT+F pipeline of Drews et al. [6] on DD).

Against **gaze-only baselines**, local optical flow results in an improvement for fixation detection on DD, where event-level fixation F_1 rises from 54 (I-VT) and 59 (I-DT) to 77 (I-VT+F) and 75 (I-DT+F). On GrW, the corresponding comparison between RF-LOCALFLOW

(FULL) and RF (GAZE-ONLY) shows consistent and statistically significant gains: event-level F1+FO F_1 increases from 56.4 to 77.1 and SACCAD from 76.2 to 80.5 (paired Wilcoxon signed-rank across folds, Holm-corrected $p \leq .006$ for both). Therefore, local flow provides useful motion context that gaze velocity alone cannot supply in head-mounted settings, particularly for producing temporally coherent fixation and saccade events.

For the **global-motion baselines**, the benefits are less clear. On DD, replacing the original globally averaged Lucas-Kanade flow with our dense gaze-centred NeuFlow v2 flow leaves event-level fixation F_1 essentially unchanged for I-VT+F (77 vs. 77) and only marginally lower for I-DT+F (72.8 vs. 75). This indicates that, for fixation and saccade detection on this dataset, the two flow formulations are effectively interchangeable. In other words, local optical flow preserves the strong fixation-detection behaviour of the Drews et al. [6] framework, but does not provide an additional advantage for these two classes. On GrW, the rule-based 2-THRESHOLDS (which already uses local flow for velocity compensation) even outperforms all other variants at the sample level for F1+FO and SACCAD, suggesting that compensated gaze velocity together with adaptive thresholds and post-processing is already close to saturation for these classes.

Overall, the answer to RQ1 is that local optical flow is *beneficial over gaze-only baselines* but *roughly equivalent to global-motion baselines* for fixation and saccade detection. For these two event types, what matters is that some form of scene-motion compensation is available, not whether it is computed globally or locally.

5.2 RQ2. Local Optical Flow for Smooth Pursuit Detection

This research question examines whether local optical flow enables smooth pursuit detection, which cannot be done reliably with gaze-only or global flow inputs. Simply adding a second velocity threshold on the locally compensated relative velocity (2-THRESHOLDS) is not sufficient; this is expected because, after motion compensation, both VOR-driven stabilisation and true smooth pursuit can produce similarly low relative velocities, and no single threshold can reliably separate them (Appendix ??). Conversely, removing local-flow features entirely (RF (GAZE-ONLY)) reaches 18.5 event-level pursuit F_1 , confirming that gaze signals alone carry too little information for pursuit detection in dynamic, head-mounted settings. When local flow is instead transformed into structural features around the gaze point, the benefit becomes substantial. RF-LOCALFLOW (FULL) reaches an event-level pursuit F_1 of 44.9, with the single-feature variants ranging from 43.6 to 46.8, compared with 27.2 for the strongest baseline (GRU) and 12.7 for ACE-DNV, which relies on visual odometry. The UNIFIED CLASSIFIER also attains one of the highest event-level weighted averages (77.7), on par with the other RF variants, showing that these pursuit gains are not obtained at the cost of the other classes. For the ablation studies, removing all flow features (RF (GAZE-ONLY)) produces the largest drop and is the only ablation that significantly reduces performance relative to the full model (paired Wilcoxon signed-rank across folds; Holm-corrected $p \leq .016$ for all three classes). In contrast, removing any single flow feature leaves performance statistically similar to the full model, i.e., none of the single-feature removals differs significantly on any class

(all Holm-corrected $p > .3$), and the small variations in event-level pursuit F_1 across these variants (43.6–46.8) lie well within the ± 6 –8 fold-level standard deviations reported in ?? We stress that smooth pursuit occurs in only 7 folds, so the redundancy family (full vs. each single-feature removal) is tested at $n = 7$; these non-significant results reflect insufficient power to resolve small per-feature effects rather than positive evidence that any feature is dispensable. Therefore, we do not interpret the ranking among single-feature variants; the nominally higher pursuit scores obtained without the relative-motion or cosine feature might reflect fold-level variance rather than a genuine effect. The effect sizes, which are less sensitive to n , point the same way: the single-feature removals yield small, inconsistent-sign rank-biserial correlations (r between -0.79 and $+0.64$), whereas removing all flow features gives a saturated effect ($r = +1.00$), the contrast the design is powered to detect. All four flow features are derived from the same dense gaze-centred flow field and encode overlapping object-relative motion cues. Therefore, the improvement in pursuit detection is best attributed to the local-flow feature set *jointly*, rather than to any single feature, and it is lost only when object-relative motion information is removed altogether. A dedicated redundancy analysis (e.g., pairwise feature correlation or permutation importance) would quantify this overlap directly and is left for future work. Dense optical flow is essential because patch-level statistics, such as foveal-peripheral flow difference and entropy, cannot be computed from sparse Lucas-Kanade flow, which partly explains why global-flow pipelines miss smooth pursuit. Additionally, GRU achieves the best PURSUIT score at the sample level (28.8), while our local-flow RF achieves the best PURSUIT score at the event level; this suggests that temporal sequence models and local-flow features are complementary, with temporal context helping frame-wise labelling and local-flow descriptors helping produce coherent pursuit events under majority-voting matching. In summary, the answer to RQ2 is yes, provided that local flow is transformed into structural features around the gaze point rather than merely used for velocity compensation. However, it is important to note that in GrW, smooth pursuit occurs only in the single Ball-Catching task, so these pursuit findings should be read as evidence from one task rather than across the full range of natural head-mounted behaviour, and broader validation remains necessary.

5.3 RQ3. Latency and Real-Time Feasibility

To address RQ3, we integrated both proposed approaches into a streaming pipeline based on MSP [3] and replayed GrW recordings as a 30 FPS input. The 2-THRESHOLDS approach inherits the runtime profile of the underlying I-VT/I-DT pipeline, plus an additional threshold comparison per sample, and is negligible in cost once optical flow is available. The richer RF-LOCALFLOW (FULL) classifier sustains 30 FPS throughput with ≈ 40 ms end-to-end latency on the tested desktop GPU setup, with optical-flow estimation dominating the runtime (≈ 30 ms) and feature extraction (≈ 3 ms) and classification (≈ 7 ms) contributing only a small fraction. A latency in the tens of milliseconds at 30 FPS is within the range typically considered acceptable for online applications of head-mounted eye tracking, such as gaze-contingent interaction and real-time visual behaviour

analysis. The breakdown also makes clear that the throughput bottleneck lies in dense optical-flow computation rather than in our features or classifier, so faster flow estimators would translate most directly into lower latency.

Our setup uses gaze and scene video only, without IMUs or other additional sensors. We used NeuFlow v2 because it provides dense optical flow with a favourable accuracy-throughput trade-off. Although NeuFlow v2 has itself reported real-time results on edge hardware [28], we evaluated our full pipeline only on a desktop GPU, and these measurements should not be read as evidence of edge readiness. Whether the approach can meet real-time constraints on mobile or wearable platforms is an open question we did not test and regard as a natural next step. On [GitHub](#), we have a video demonstration of the live pipeline.

5.4 Limitations

This study has several limitations that constrain how far the results generalise beyond the tested datasets and tasks. First, GrW’s smooth pursuit labels come almost entirely from the “Ball Catching” task, so the reported pursuit gains may not transfer to slower targets, cluttered backgrounds, or different viewing distances; this partly reflects how few head-mounted datasets offer fine-grained pursuit annotations in the first place. Second, our real-time evaluation used a desktop GPU, which we report for transparency. The approach itself is not tied to this hardware, since optical flow estimation dominates the measured runtime, not our features or classifier; a faster or edge-optimised estimator is the clearest path to lower-power deployment, though this remains to be measured directly on mobile or wearable platforms. Third, design choices such as the fixed patch geometry, feature definitions, and post-processing constraints introduce hyperparameters whose optimal values may be dataset- and device-dependent. Finally, pursuit gains are stronger at the event level than the sample level, a pattern that helps applications needing coherent event segmentation but leaves room to improve frame-wise consistency, for example, by combining local-flow features with lightweight temporal modelling.

5.5 Ethical Considerations & Responsible Innovation

Eye-tracking, especially real-time inference, can introduce privacy and misuse risks, since gaze behaviour may reveal sensitive information and enable covert monitoring if deployed without strong safeguards (e.g., informed consent, data minimisation, purpose limitation). Additionally, reliability may vary across participants, environments, and devices, but this work does not provide a systematic analysis to quantify such differences.

6 Conclusion & Future Work

In this study, we built on the approach of Drews et al. [6] by incorporating local optical flow around the gaze point to improve smooth pursuit detection in head-mounted eye tracking, while preserving their framework’s fixation-detection performance. 2-THRESHOLDS, a direct rule-based extension using I-VT and I-DT, performed well for FI+FO and SACCADE but failed for PURSUIT, showing that rule-based criteria alone are insufficient for pursuit detection in dynamic head-mounted scenes. On the GrW “Ball Catching” task, the only

source of pursuit labels in the dataset, our unified RF with local-flow features achieved the best event-level PURSUIT performance ($F_1 = 44.9$, up to 46.8 in the best variant) while remaining competitive overall (weighted average $F_1 = 77.7$), showing that local optical flow benefits pursuit detection without sacrificing performance on the other classes. An implementation of our approach is available in the OpenGazeLab toolkit [19], and our codebase is publicly available on [GitHub](#). Future work should evaluate generalisation beyond GrW, assess real-time deployment on edge devices such as smartphones paired with head-mounted headsets, and explore lightweight temporal modelling to improve sample-level consistency and better handle short or fragmented pursuit events.

Acknowledgments

We thank Benjamin Stahr (University of Oldenburg) for his thesis, which helped inspire and motivate this work. This work was funded by the Federal Ministry of Research, Technology and Space (BMFTR) under grant number 16IW23002 (No-IDLE) and grant number 16IW24006 (NoIDLEChatGPT), the Lower Saxony Ministry of Science and Culture (MWK) in the zukunft.niedersachsen program, and the Endowed Chair of AAI at the University of Oldenburg.

References

- [1] Rochelle Ackersley et al. 2011. The interaction of visual, vestibular and extra-retinal mechanisms in the control of head and gaze during head-free pursuit. *The Journal of Physiology* 589, 7 (2011), 1627–1642. doi:10.1113/jphysiol.2010.199471
- [2] Richard Andersson et al. 2017. One Algorithm to Rule Them All? An Evaluation and Discussion of Ten Eye Movement Event-Detection Algorithms. *Behavior Research Methods* 49, 2 (April 2017), 616–637. doi:10.3758/s13428-016-0738-9
- [3] Michael Barz et al. 2021. Multisensor-Pipeline: A Lightweight, Flexible, and Extensible Framework for Building Multimodal-Multisensor Interfaces. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. ACM, Montreal QC Canada, 13–18. doi:10.1145/3461615.3485432
- [4] Asim H. Dar et al. 2021. REMoDNaV: Robust Eye-Movement Classification for Dynamic Stimulation. *Behavior Research Methods* 53, 1 (Feb. 2021), 399–414. doi:10.3758/s13428-020-01428-x
- [5] M. Dorr et al. 2010. Variability of Eye Movements When Viewing Dynamic Natural Scenes. *Journal of Vision* 10, 10 (Aug. 2010), 28–28. doi:10.1167/10.10.28
- [6] Michael Drews et al. 2024. Strategies for Enhancing Automatic Fixation Detection in Head-Mounted Eye Tracking. *Behavior Research Methods* 56, 6 (April 2024), 6276–6298. doi:10.3758/s13428-024-02360-0
- [7] Carlos Elmadjian et al. 2022. Online Eye-Movement Classification with Temporal Convolutional Networks. *Behavior Research Methods* 55, 7 (Oct. 2022), 3602–3620. doi:10.3758/s13428-022-01978-2
- [8] Xiaoxue Fu et al. 2024. Implementing Mobile Eye Tracking in Psychological Research: A Practical Guide. *Behavior Research Methods* 56, 8 (Aug. 2024), 8269–8288. doi:10.3758/s13428-024-02473-6
- [9] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley]. <https://www.jstor.org/stable/4615733>
- [10] Kenneth Holmqvist et al. 2011. Eye Tracking: A Comprehensive Guide to Methods and Measures. *EBL* (2011). <https://global.oup.com/academic/product/eye-tracking-9780199697083>
- [11] Sabrina Hoppe et al. 2016. End-to-End Eye Movement Detection Using Convolutional Neural Networks. doi:10.48550/arXiv.1609.02452
- [12] László Kopácsi et al. 2026. Gaze-Based Menu Navigation in Virtual Reality: A Comparative Study of Layouts and Interaction Techniques. In *Human-Computer Interaction – INTERACT 2025*. Springer Nature Switzerland, Cham, 520–543. doi:10.1007/978-3-032-04999-5_31
- [13] Rakshit Kothari et al. 2020. Gaze-in-Wild: A Dataset for Studying Eye and Head Coordination in Everyday Activities. *Scientific Reports* 10, 1 (Feb. 2020), 2539. doi:10.1038/s41598-020-59251-5
- [14] Linnéa Larsson et al. 2016. Head movement compensation and multi-modal event detection in eye-tracking data for unconstrained head movements. *Journal of Neuroscience Methods* 274 (Dec. 2016), 13–26. doi:10.1016/j.jneumeth.2016.09.005
- [15] Bruce D. Lucas et al. 1981. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2 (IJCAI'81)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 674–679. <https://www.ijcai.org/Proceedings/81-2/Papers/017.pdf>
- [16] Abdulrahman Mohamed Selim et al. 2024. A Review of Machine Learning in Scanpath Analysis for Passive Gaze-Based Interaction. *Frontiers in Artificial Intelligence* 7 (June 2024), 1391745. doi:10.3389/frai.2024.1391745
- [17] Ashkan Nejad et al. 2024. ACE-DNV: Automatic Classification of Gaze Events in Dynamic Natural Viewing. *Behavior Research Methods* 56, 4 (March 2024), 3300–3314. doi:10.3758/s13428-024-02358-8
- [18] Sharon Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger (Eds.). 2019. *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions*. Association for Computing Machinery and Morgan & Claypool. <https://dl.acm.org/doi/book/10.1145/3233795>
- [19] Khue Minh Pham et al. 2026. OpenGazeLab: An Interactive Toolkit for Gaze Analysis and Event Detection. In *Proceedings of the 2026 Symposium on Eye Tracking Research and Applications (ETRA '26)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3797246.3805855
- [20] Alexander Plopski et al. 2023. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *Comput. Surveys* 55, 3 (March 2023), 1–39. doi:10.1145/3491207
- [21] Dario D. Salvucci et al. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00*. ACM Press, Palm Beach Gardens, Florida, United States, 71–78. doi:10.1145/355017.355028
- [22] Abraham Savitzky and M. J. E. Golay. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36, 8 (July 1964), 1627–1639. doi:10.1021/ac60214a047
- [23] Mikhail Startsev et al. 2019. 1D CNN with BLSTM for Automated Classification of Fixations, Saccades, and Smooth Pursuits. *Behavior Research Methods* 51, 2 (April 2019), 556–572. doi:10.3758/s13428-018-1144-2
- [24] Mikhail Startsev et al. 2023. Evaluating Eye Movement Event Detection: A Review of the State of the Art. *Behavior Research Methods* 55, 4 (June 2023), 1653–1714. doi:10.3758/s13428-021-01763-7
- [25] Julian Steil et al. 2018. Fixation Detection for Head-Mounted Eye Tracking Based on Visual Similarity of Gaze Targets. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, Warsaw Poland, 1–9. doi:10.1145/3204493.3204538
- [26] Marc Tonsen et al. 2020. A High-Level Description and Performance Evaluation of Pupil Invisible. *arXiv:2009.00508*
- [27] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. Publisher: [International Biometric Society, Wiley]. doi:10.2307/3001968
- [28] Zhiyong Zhang et al. 2025. NeuFlow v2: Push High-Efficiency Optical Flow To the Limit. doi:10.48550/arXiv.2408.10161