A pose detection framework, validated through a bicycle simulator experiment

Csaba Hajdu*, Pál Ágoston Sándor, Katalin Kovács, Amira Hammami and Attila Borsos

Széchenyi István University
Egyetem tér 1, H-9026 Győr, Hungary
hajdu.csaba@ga.sze.hu*, sandor.agoston.pal@ga.sze.hu, kovacsk@sze.hu, hammami.amira@sze.hu, borsosa@sze.hu
*Corresponding author

Abstract: *Cyclists constitute one of the most Vulnerable Road Users (VRU) groups, yet the majority of traffic-analysis studies, whether simulation-based or empirical, continue to focus on vehicle traffic simulation and, to a lesser extent, pedestrian behavior. Consequently, the interaction patterns that arise when cyclists share the roadway with motor vehicles remain comparatively underexamined, even though miscommunication of intent between drivers and cyclists is a leading contributor to urban collisions. To address these shortcomings, the present study introduces an integrated IT hardware-in-the-loop (HiL) simulation framework, that embeds an instrumented bicycle, equipped with inertial sensors, into realistic game engine-based traffic scenarios. The framework and its associated measurement ecosystem (vehicle dynamics signals, eye-tracking, physiological signals and body posture) establish a reproducible foundation for cycling-oriented perception research and support the development of safety-critical driver-assistance and autonomous driving functions that recognize cyclists as primary traffic participants rather than peripheral actors. In general, this paper provides a flexible and scalable test environment, at the intersection of cyclist traffic safety, human-machine interaction and infrastructure design.*

Keywords: pose detection; bicycle simulator; gaze behavior; hardware-in-the loop; behavior detection

1 Introduction

Cyclists are among the most vulnerable road users (VRU), yet traffic-analysis studies, whether simulation-based or empirical, still focus mainly on vehicles and, to a lesser extent, on pedestrians. The interaction between cyclists and motor vehicles therefore remains underexamined, although miscommunication of intent between drivers and cyclists is a leading contributor to urban collisions.

Vision-based human-pose estimation and gaze analysis have proved invaluable for modeling pedestrian intent, but bicycle-centric investigations are sparse for two reasons: (i) the rider's body is partly occluded by the bicycle frame, reducing the reliability of generic pose detectors; and (ii) public benchmarks seldom contain cyclist behaviors, such as hand signals, shoulder checks, or gaze cues, from the rider's point of view, which hampers benchmarking, reproducibility and algorithmic progress.

To address these shortcomings, the present study introduces an integrated *hardware-in-the-loop (HiL)* simulation framework that embeds an instrumented bicycle, equipped with inertial sensors, into realistic game engine-based traffic scenarios. A head-mounted eye-tracker records gaze direction, synchronized cameras capture egocentric and exocentric video, and multiple participants negotiate identical scripted scenarios, enabling controlled comparison across riders.

The principal contributions are as follows:

- Cyclist-centric perception framework: a HiL platform fusing inertial measurements, multi-view video, and gaze tracking to capture behavioral signals in real time.
- Graph-based scenario modeling: traffic entities and events are expressed as reusable graph templates, enabling systematic variation of environmental factors and rapid prototyping of test cases.
- Curated multi-modal dataset: synchronized sensor streams, calibrated imagery, pose skeletons, and gaze annotations are consolidated into a structured dataset that underpins the evaluations reported herein.

Together, the framework and the dataset provide a reproducible foundation for cycling-oriented perception research and for safety functions that treat cyclists as primary traffic participants.

2 Related work

Research on understanding traffic scenarios has historically prioritized motor vehicle dynamics and, more recently, pedestrian analysis; consequently, cyclists have received comparatively limited attention despite being one of the most vulnerable road users (VRU) groups. This section reviews the state of the art in bicyclist-centric perception, with a focus on:

1. Vision-based cyclist detection and behavior classification
2. Cyclist-specific pose estimation
3. Gaze- and head-based intent inference
4. Datasets that enable empirical progress

2.1 Vision-Based Cyclist Detection and Behavior Classification

Object detection backbones such as YOLO (You Look Only Once) [1] and derivatives YOLOv12 [2], EfficientNet [3], Faster-RCNN [4] (Faster Region Based Convolutional Neural Networks) and similarly Mask-RCNN [5] routinely exceed 90% mean average precision for vehicle classes. However, their performance in detecting cyclists deteriorates when these networks are trained on canonical traffic benchmarks (e.g., KITTI [6], nuScenes [7]) that underrepresent the appearances and viewing angles of cyclists [8]. To mitigate this imbalance, 3DArticCyclists have been proposed [8], a synthetic rendering pipeline that generates photorealistic cyclist images with controllable viewpoints and articulated poses, increasing downstream detection accuracy by up to 7% when used for data augmentation.

Beyond localization, several studies address recognition of cyclist behavior. BeST-DAD [9] was introduced with a convolutional auto-encoder that flags trajectory anomalies (e.g., sudden swerves, erratic deceleration) in GPS and video streams, outperforming PCA-based baselines on a 22-h naturalistic-cycling corpus. At a higher semantic level, another solution focused on group cycling [10], classifying formations (single file, side-by-side) and collective maneuvers via a 3D-CNN fed with short video bursts. These works show a shift from mere detection towards semantic understanding of cyclist actions.

2.2 Pose Estimation for Cyclist Motion Understanding

Human-pose estimation provides fine-grained kinematic cues indispensable for early-intent prediction. While open-domain frameworks such as OpenPose, HRNet, BlazePose detect two-dimensional (2D) key points in real time, their reliability degrades for riders whose lower limbs are partially occluded by the bicycle frame. It has been demonstrated that a 2D pose sequence, embedded via an LSTM, anticipates cyclist start/turn intentions 0.8 seconds earlier than motion-only baselines [11]. More recently, multi-view stereo have been exploited to recover three-dimensional (3D) rider pose, enabling recognition of subtle pre-turn preparations (e.g., lean torso) at intersections [12]. However, collecting large-scale pose ground truth for cyclists remains an open challenge, motivating reliance on synthetic augmentation and weak supervision [8].

2.3 Gaze and Head Orientation as Intent Cues

Visual attention strongly modulates VRU decision-making. An in-the-wild eye-contact dataset and trained a two-stream CNN to classify whether a pedestrian or cyclist is looking at the ego-vehicle, with a neural network model, that achieved 79% F1 score at distances up to 25 meters [13]. Extending this idea, a spatio-temporal graph of skeletal keypoints has been incorporated to detect eye contact at 40 m, highlighting the benefit of body context [14]. Moving beyond binary contact, the recently introduced *PEdestrian Gaze Object* (PEGO) task and a

transformer that predicts which scene object attracts the VRU's gaze without explicit eye images [15]. For cyclists, recognizing over-the-shoulder "head-checks" is particularly informative for turning intent, but large-scale annotations are still lacking; preliminary work adapts pedestrian gaze schemes to cyclist head pose with encouraging results [12].

2.4 Datasets Supporting Cyclist Behavior Research

The datasets used globally for behavior detection are listed in Table 1, focusing on traffic scenes and particularly pedestrians. While *Joint Attention in Autonomous Driving* (JAAD) and *Pedestrian Intention Estimation* (PIE) datasets pioneered gaze annotations, they contain few cyclists. *Cyclist Arm Signal Recognition* (CASR) partially fills this gap, but remains relatively small-scale [11]. Synthetic resources generated from simulations therefore play an essential role in balancing data and diversifying rare (and potentially dangerous) maneuvers, traffic situations, and otherwise dangerous situations (e.g., accidents). Many dataset images are generated with the help of high-end graphics, using Unreal Engine for example, like in CARLA [16]. In most cases, these datasets focus on the perspective of one actor, such as in the case RealStreet and SynthStreet [17]. Synthetic data set generation can be further augmented with different perspective views, capturing real images and geographic data (using OpenStreetMap - OSM), such as in the case of RealStreet derivatives[18]. Such data sets are used to bridge the gap in traffic signal control and traffic situations, where simulations of such scenarios are mandatory to ensure proper safety. Such data sets are available publicly, for example SynTraC - Synthetic Dataset for Traffic Signal Control from Traffic Monitoring Cameras[18].

2.5 Synthesis and Research Gaps

The literature demonstrates tangible progress towards cyclist-aware perception. However, two main limitations persist:

1. Data scarcity, especially for gaze and high-fidelity pose under occlusion
2. Domain robustness, as models trained in one city frequently underperform in novel lighting or cultural contexts

Ongoing work on cross-domain adaptation and sensor fusion (camera & radar) shows promise, yet comprehensive benchmarks for cyclist intent prediction akin to PIE are still outstanding.

Table 1
Datasets supporting behavior analysis

Dataset	Size/Modality	Key Annotations	Ref.
JAAD	346 dash-cam clips	Pedestrian gaze, attention, actions	[19]
PIE	6+ hours of urban driving video	Frame-level intention labels (cross/not cross), gaze cues	[20]
CASR	Real-world & Web videos	2D pose sequences, cyclist & pedestrian intent classes	[11]
3DArticCyclists	Photorealistic synthetic imagery	Articulated cyclist poses from diverse viewpoints	[8]
BeST-DAD	Large-scale GPS + video traces	Anomaly labels for cyclist safety studies	[9]
SynTraC	Large Photorealistic synthetic dataset	Traffic signal control	[18]
RealStreet & SynthStreet	Large Photorealistic synthetic dataset	Generated traffic situations	[17]
RealStreet derivatives	Large augmented photorealistic synthetic dataset	Generated traffic situations	[18]

3 Methodology

The main goal of this research is to simulate traffic situations from a cyclist point of view. The research focuses on detecting reaction to dangerous situations by investigating cognitive reflex-action to these situations.

The simulation system itself is a construction of a bicycle, with the proper setup to measure the state of the bicycle and the cyclist. From a cyber-physical sense [21], the bicycle acts as the peripheral device that can interact with the simulation through the user.

- **Cyclist**: the cyclist rides the bicycle, and the *eye-movement* and the heartbeat is continuously measured. These factors contribute to attention in traffic situations.
  - *Heartbeat*: the measured heartbeat of the cyclist.
  - *Eye movement*: the eye movement of the cyclist is investigated, including the saccadic motion, the time length of keeping focus on a certain point, and blinking period.

- **Bicycle**: the bicycle command (a vector value: steering, brake, acceleration $\dot{\phi}, \vec{F}_{brake}, a$) is measured, with the roll θ of the bicycle. So, the typical state description for the bicycle $[x, \dot{x}, \theta, \dot{\theta}]$ is used.
  - *Roll* θ: the instantaneous roll rotation of the bicycle, relative to the frame of the bicycle (changed by the steering value).
  - *Steering* $\dot{\phi}$: steering angle of the bicycle governed by the cyclist.
  - *Brake* $\vec{F}_{brake}$: brake induced by the driver, to decrease the velocity of the bicycle.
  - *Acceleration* a: acceleration induced by the driver to increase the velocity of the bicycle.

The simulation system outputs the behavior, the pose (detected kinematic tree), and the 3D-position of the cyclist in the HiL simulation scene. The simulation also outputs different objects related to the simulation, including vehicles and objects (obstacles). The input and output of the simulation are depicted in Figure 1. The simulation contains the scene (modeled after the real street in Győr) and the traffic flow simulation.

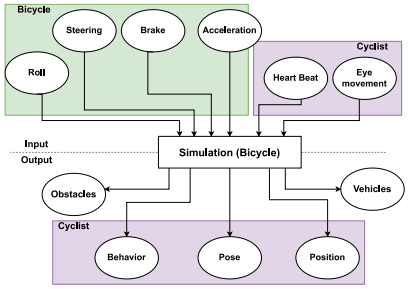


Figure 1
Input-output model of the system

The bicycle and the cyclist are observed in a controlled environment (depicted in Figure 2) with a camera-based approach. While the gaze is detected on the cyclist, a depth camera in front of the subject observes the cyclist and its movement (e.g., to detect the pose and movement). The cyclist rides the bicycle in the virtual reality (HiL simulation) containing actors and obstacles. Cyclists are continuously measured in varied traffic flow of the same traffic scene [22].

The simulation itself is an Unreal-based simulator, containing the digital replica of a street in Győr, Hungary with pedestrians, vehicles, and sidewalks. The simulation is depicted in Figure 3, indicating the view of the simulator, the eye movement, gaze direction and the world itself.

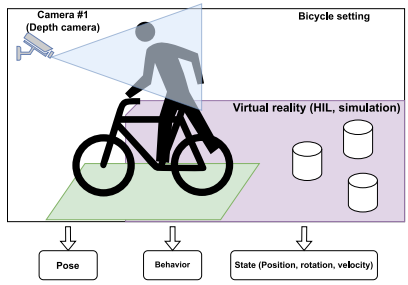
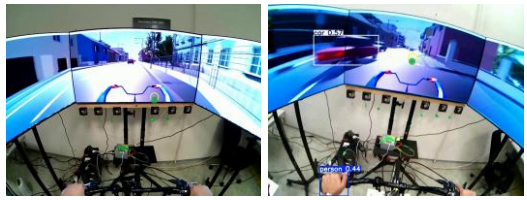


Figure 2
Observation of the cyclist in the environment

In conclusion, for the HiL loop, the poses, the bounding boxes of objects and inference time can be easily extracted as a measure to verify the dataset.



(a) World simulation (b) Object detection

Figure 3
Simulated world from the cyclist viewpoint

3.1 Hypergraph-based formalism

Compositionally, the subject is a bicycle ridden by a cyclist: the cyclist is described by the usual human kinematic tree, whereas the bicycle contains only the minimal set of kinematic components (Figure 4). Both are described by a tree of keypoints with their instantaneous position and rotation; MS COCO, for example, defines 17 keypoints.

The measurements and their relations to the dependent variables are stored in a hypergraph-based formalism, which concisely organizes the experiments, the results, and how each metric relates to the cyclist; its main ideas are the reuse of, and the reference between, fragments of semantic information. The description was inspired by cognitive infocommunications [23], focusing on the actuator-sensor interface of cognitive agents. Listing 1 shows the description of a single cyclist (all quantities are measured during the simulation and exported as CSV): the *gaze positions*, *blinks*, *saccades*, *fixations* (with their duration), and the *3D eye pose* (optical axis, eyeball center, and pupil diameter), where the eye is defined once and reused for the left and right eyes. The world objects perceived by the cyclist carry bounding boxes, positions, trajectories, and a visibility value.

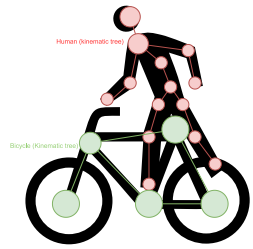


Figure 4
Kinematic composition of the cyclist and the bicycle peripheral

Listing 1
Bicycle Behavior Dataset Schema

```
[bicycle description]
bicycle {
  person name ["AS", "AW", "AV", "AW", "AX"]
  eye movement {
    gaze positions {   timestamp<int>      pos [0.0, 0.0]   }
    blinks { id<int>, start<int>,  end<int>,   duration<int>  }
    saccades { id<int>,  start<int>,    end<int>,  duration<int>, amplitude [0.0,
  0.0]  # polar coordinates
      velocity [0.0, 0.0]   # mean & peak  }
    fixations { id<int>, start<int>, end<int>, duration<int>, fixation [0.0, 0.0]  }
    eye pose 3d { timestamp<int>,  eye { pupil diameter [0.0, 0.0], eyeball center
  [0.0, 0.0, 0.0]
        optical axis [0.0, 0.0, 0.0]    }
      left:  |> -> eye {}
      right: |> -> eye {}
    }
  }
  world {
    object { id<int>, name<string>, duration<int>, bounding box [0.0, 0.0, 0.0,
  0.0], world_position [0.0, 0.0, 0.0]
      trajectory { timestamp<int>, position [0.0, 0.0, 0.0], velocity [0.0, 0.0,
  0.0]  }
      visibility<float>
    }
  }

  # file bindings (filenames) and loader hyperedges (@load_data) omitted
}
```

3.2 Computer vision framework

In addition to the embedded device that captures the pose and objects of the measurement space, multiple utility programs (with UI) were created, all aiding the research (using the Ultralytics package):

- ***Object detection utility program*:** object detection utility program, to detect objects and their corresponding bounding boxes. The program was useful in detecting objects in the simulated view.

- ***Pose detection utility program*:** to investigate videos and detected poses separately from the research in question. The target was to detect the poses of cyclists.

In addition to utility programs to test neural network models, a toolchain was used. The background object detection model was YOLOv11[24] capable of detecting poses trained on the COCO data set. Alternatively, *vision transformer pose detection* (ViTPose) [25] was used with *Real-Time Detection Transformer* (RT-DETR)[26] object detection ecosystem.

The architecture of the toolchain indicated by the inputs, is shown in Figure 5.

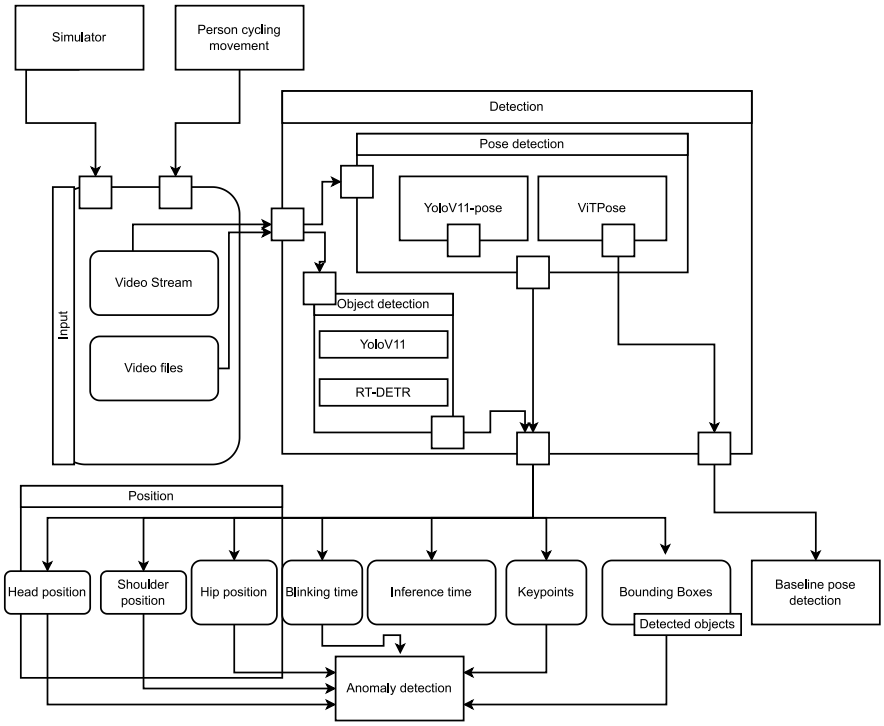


Figure 5
Architecture of inputs and outputs

3.2.1 Neural network inference

The following target objectives were measured as part of the framework:

- **Pose detection**: detection of poses of the cyclists, which is important to prepare behavior detection as well to infer anomalies in motion. However, driving styles can differ, so pose detection considers distribution for each cyclist. For pose detection, ViTPose was used as a baseline to further employ precision calculation.

- **Object detection**: detection of objects is also important: while the camera detects only the cyclists (i.e., a person) the simulation itself contains object that can induce anomalous or avoidant behavior.
- **Behavior detection**: while this article concentrates on the framework to ground further work on annotating and processing sequential data.

The benchmarking framework is based on PyTorch, with extensions to Ultralytics YOLO and HuggingFace transformers for ViTPose.

3.3 Identifying behavior pattern

Based on the cycling and simulation records, the following measurements can be taken to classify a behavior that reacts to danger:

- ***Hip position estimation*:** as the COCO pose data set identifies hips with two key points (namely the 11th and 12th), the change in the middle point can be used to analyze hip movement detection. As a reference, the shoulder mid-point can be used to define the hip more precisely (5th and 6th keypoint). Given these keypoints, the center can be calculated as $\vec{c}_{shoulder} := \frac{\vec{p}_5+\vec{p}_6}{2}$ and $\vec{c}_{hip} := \frac{\vec{p}_{11}+\vec{p}_{12}}{2}$, therefore, the center change is:

$$\Delta_{hip} := \sqrt{\left(\frac{\vec{c}_{shoulder}+\vec{c}_{hip}}{2}\right)^2} \tag{1}$$

- ***Blinking duration*:** blinking duration is measured by the embedded device, along with the gaze direction (in polar coordinates). Increased blinking duration indicates increasing focus, therefore the event of a dangerous situation. A longer duration is everything out of the normal blinking range. In the article, it is represented as $T_{blink}(t)$, to indicate the period time in seconds in any time point.
- ***Sudden head movement*:** a sudden head movement can also be considered as an event of dangerous situation. With a similar approach, the head tilting can be captured by calculating the center of the head movement, namely the center of the first (0th index) and the second (1st index) keypoints. Similarly, to Equation (1), the head center Δ_{head} is defined as:

$$\Delta_{head} := \sqrt{\left(\frac{\vec{c}_0+\vec{c}_1}{2}\right)^2} \tag{2}$$

From the exteroceptive perspective, when an object is close to the person, an anomaly can occur. For this, the distance of a detected object is estimated. The exact distance of an object given the camera picture cannot be determined, only an estimate can be determined even given the camera parameters. Given the camera parameters, the following equation can be used for a distance estimate (h_{height}

indicates the real height of an object, $\hat{h}_{\text{bounding_box}}$ indicates the height of the detected bounding box):

$$\Delta_{object} := \frac{f_{focus} \times h_{object}}{\hat{h}_{\text{bounding_box}}} \tag{3}$$

3.3.1 Modal logic formalism

Based on the measurements, formal modal logic specifications can describe identified anomalous behavior. Modal logic helps to capture the order and time-dependence of behavior, as is expected in real-time scenarios. The following scenarios are described using modal logic, computation tree logic (CTL). If $A(t)$ logical prefix indicates an anomaly occurring (i.e., $\Diamond A(t)$:

- When Δ_{hip} the center hip position (Equation (1)) change absolute value reaches a threshold $theta$, there is an anomaly:

  $$\forall \blacksquare \left(\left(\Delta_{hip}(t) \leq \sigma_{hip} \right) \rightarrow \neg A \right)$$

- When the blinking period increases, there is an anomaly event, likely a traffic incident: $\forall \blacksquare \left(F_{blink}(t) \rightarrow \Diamond E(t) \right)$
- Head movement induces a similar anomaly as hip movement: $\forall \blacksquare \left((\Delta_{head}(t) \leq \sigma_{head}) \rightarrow \neg A \right)$
- Shoulder movement induces anomaly, similarly to head movement and hip movement: $\forall \blacksquare \left((\Delta_{shoulder}(t) \leq \sigma_{head}) \rightarrow \neg A \right)$
- Close objects (e.g., cars, pedestrians, static obstacles) can indicate a dangerous event (each with its own τ threshold): $\forall \blacksquare \left(\left(\Delta_{car}(t) \leq \tau_{car} \vee \Delta_{pedestrian}(t) \leq \tau_{pedestrian} \vee \Delta_{static}(t) \leq \tau_{static} \right) \rightarrow \neg A \right)$

3.4 Experimental setup

The experiment of observing and inferring on cyclists, itself can be described with the following steps:

1. The cyclist starts riding the bicycle mock with the eye camera worn. The simulated scene starts in the simulation with the varied traffic flow.
2. The cyclist rides the bicycle, until a point is reached in the simulation. The movement in the simulation depends on the ride pace, but typically lasts around 40 seconds.
3. The experiment is repeated arbitrarily. The captured videos and time-series data is then used as the input for the statistical evaluation.
4. Besides visual data, the accelerometer feedback, the eye gaze is measured.

The statistical computations for the experiments, was run on 3 separate machines:

- A Laptop computer equipped with NVIDIA Geforce 3070 RTX, AMD Ryzen 9 5900HX, 32 GB Memory.
- A desktop computer equipped with NVIDIA Geforce 2070 Super GTX, Ryzen 7 3700X, 8-core, 32 GB memory.
- A HPC system (DKF KIFÜ Komondor) equipped with multiple nodes of NVIDIA A100, capable of running parallel simulations.

The inclusion of HPC systems allows massively parallel benchmarks focusing on behavior analysis.

4 Results

The results include a large-scale dataset, each containing a large set of videos for each cyclist from different perspectives and measurements (including gaze movement, accelerometer feedback). Both videos contain visualization of the gaze of the cyclist's eye, the eye movement (including saccades), and visualization of the pose of the cyclist (depicted in Figure 6). Pose detection and visualization were two-fold:

1. An embedded device, which recorded the cyclist, was capable of inferring the pose using MediaPipe framework [27] used as an overlay on the resulting video.
2. A separate video analysis toolchain was applied to detect poses and match time-series measurements in the output videos, generated by the HiL system.



(a) Cyclist 1 (b) Cyclist 2 (c) Cyclist 3

Figure 6
Selection of bicyclists from the dataset (pose view)

The data set includes the following attributes (summarized in Table 2) and contains measurements from 50 participants. This includes the videos of the two perspectives and the rendered view of the simulation.

Table 2
Summary of HiL Simulation Dataset Attributes

Metric	Count
Participants	50
Total Measurements	598
Cyclist Videos	528
Simulator Videos (Cyclist POV)	548

4.1 Inference results

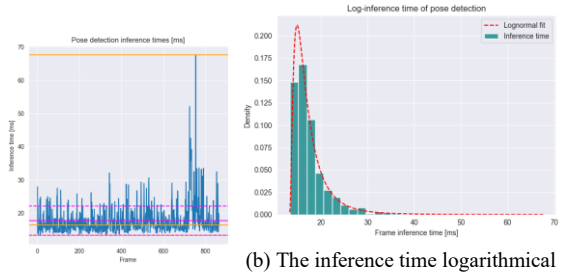
Keypoint detection was executed on the experimental setup described in Section 3.4. Overall, inference times show a right-skewed distribution, i.e., most frames are processed close to a typical value while a smaller portion exhibits slower processing due to occasional system-level effects (I/O, scheduling, background load). In the laptop setup, typical inference times were below 20 ms, while rare outliers reached up to approximately 60 ms.

The quantitative performance results are summarized in Table 3, comparing preprocessing, inference, and postprocessing stages across the tested devices and YOLO11 model sizes. Figure 7 illustrates the inference time over consecutive frames and its logarithmic representation, which highlights the skewness and supports the use of log-scale visualization for latency analysis.

In addition to single-run performance, the major advantage of the HPC environment is throughput: the ability to process large volumes of recordings by parallelizing video batches and distributing them across compute resources. This makes the workflow practical when the measurement volume scales to hundreds of recordings.

Table 3
Inference performance of YOLO11 models on different hardware

		Mean Time [ms]			Median
Model	**Device**	**Pre**	**Infer**	**Post**	**Infer [ms]**
YOLO11 (Small)	Laptop	1.83	17.65	3.12	16.39
YOLO11 (Medium)	Laptop	1.54	18.81	3.13	18.47
YOLO11 (Nano)	Desktop	1.10	6.10	1.10	6.05
YOLO11 (Medium)	Desktop	1.21	10.73	1.22	10.22

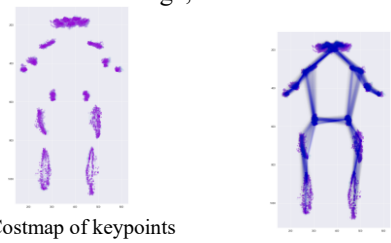


(a) Pose inference time on consecutive frames

(b) The inference time logarithmical scale still shows skewed inference times

Figure 7
Inference time of the pose detection

The pose detection induced a costmap of most possible positions of each keypoint (depicted in Figure 8). It is visible, that legs, shoulders are moving frequently, indicating hip movement and head movement more of a point of interest regarding detecting anomalies. From the ghost view of the kinematic skeleton (Figure 8b), it is visible that hip movement is not as stable or in a certain field, as the limbs: limbs are constantly moving in a certain range, even when an anomaly is occurring

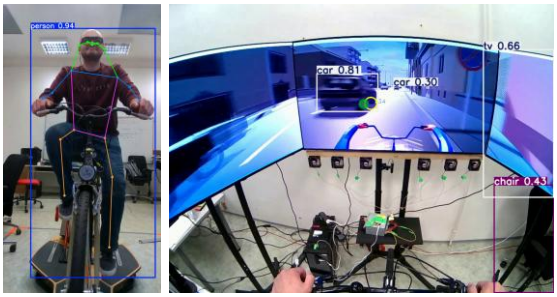


(a) Costmap of keypoints through time

(b) Skeleton indicating RoI

Figure 8
Pose detection costmap indicating outliers on a single person poses

As mentioned in Section 3.2 and Section 3.3.1, the framework outputs the anomalous frames based on the modal logic, defined in Section 3.3.1. Based on that, the frames were identified whether it was long blinking duration, sudden hip movement or head movement. Using the anomaly detection given the logic previously defined, currently in a manual way, in the most case a close car indicated an anomaly. An example for anomaly detection is depicted on Figure 9.



(a) Anomalous pose (b) Object detection in simulator indicating a dangerous situation

Figure 9
Anomaly detection on the video input

4.2 Statistical analysis

An exploratory statistical analysis was conducted to characterize three behavioral proxies extracted from the recordings: (i) center-hip displacement, (ii) center-head displacement, and (iii) blink duration. The goal of this analysis is twofold: (1) to provide compact quantitative descriptors for each proxy, and (2) to prepare subsequent probabilistic inference (e.g., Bayesian modeling) in future work.

The selected behavior attributes used for further analysis are:

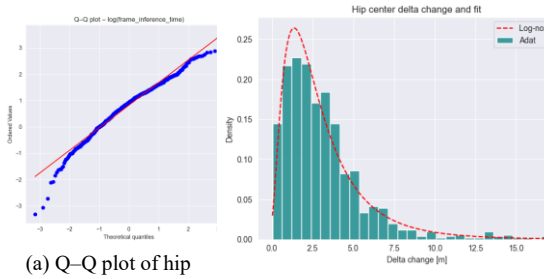
- **Center hip position change**: the displacement of the hip center, computed as the barycentric position of multiple hip-related keypoints and evaluated as a frame-to-frame change.
- **Center head position change**: the displacement of the head center, computed as the aggregate mean of head-related keypoints and evaluated as a frame-to-frame change.
- **Blinking time**: blink duration derived from the eye-related time-series measurements.

Across the evaluated samples, these signals exhibit non-negative magnitudes and visibly right-skewed empirical distributions. Therefore, a log-normal model was considered as a compact parametric approximation. Fit quality was assessed using Q–Q plots and a Kolmogorov–Smirnov (K–S) test at a significance level of $\alpha = 0.05$ (exploratory). Table 4 reports summary statistics (mean and confidence interval bounds) of the fitted distributions, while Figure 10–Figure 12 provide the corresponding Q–Q plots and histograms.

Table 4
Summary statistics of fitted distributions for behavioral attributes

			Confidence Interval	
Attribute	**Fit Distribution**	**Mean**	**Min**	**Max**
Center-hip change	Log-normal (right-skew)	3.15	0.35	10.29
Center-head change	Log-normal (right-skew)	2.86	0.27	10.83
Eye blinking	Log-normal (right-skew)	266.43	159.43	373.43

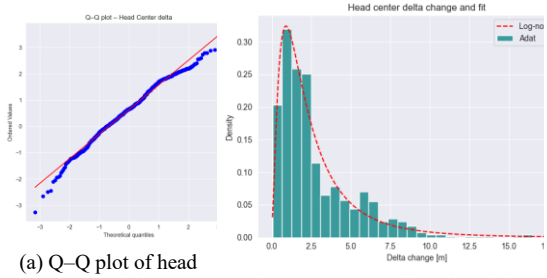
The *center hip position distribution* can be used as a proxy for hip stability. The displacement magnitudes follow a right-skewed distribution and are reasonably approximated by a log-normal model (Figure 10).



(a) Q–Q plot of hip center delta (b) Histogram and fitted model

Figure 10
Distribution of the hip center position change (delta)

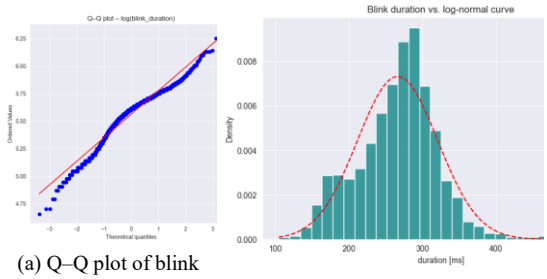
Similarly, head-center displacement magnitudes show a right-skewed distribution and can be approximated by a log-normal model (Figure 11).



(a) Q–Q plot of head center delta (b) Histogram and fitted model

Figure 11
Distribution of the head center position change (delta)

Blink durations were derived from the auxiliary eye-related time-series measurements. The empirical distribution of blink duration is also right-skewed and can be compactly represented by a log-normal model (Figure 12).



(a) Q–Q plot of blink duration

(b) Histogram and fitted model

Figure 12
Distribution of blink duration

Finally, exploratory correlation analysis was performed on log-transformed movement magnitudes to reduce the effect of extreme values. Weak-to-moderate associations were observed in selected segments (e.g., between hip and head movement, and between blink duration and hip movement); however, these findings are preliminary and should be validated on fully synchronized recordings and a larger sample in future work.

5 Discussion

The simulator and its measurement ecosystem (vehicle dynamics, eye-tracking, physiological signals, and body posture) combine laboratory-level experimental control with the ecological relevance of real traffic situations. Cyclist behavior can thus be studied risk-free, yet repeatably and comparably, while time-synchronized multimodal acquisition supports the joint analysis of attention allocation, stress-related responses and movement patterns.

The applications to date show three promising use cases. *Simulator validation*: comparing field- and simulator-based performance indicators with perceived realism measures (e.g., PLR-type indicators) supports simulator fine-tuning and bounds the interpretation of experimental outcomes. *Infrastructure and interaction design*: in automated vehicle–cyclist scenarios, eye-movement metrics and heatmaps consistently indicate that infrastructure type shapes visual strategies and situation awareness, so the simulator suits the comparison of road cross-section alternatives. *Automated safety evaluation*: pose- and video-based anomaly detection, extended by runtime monitoring, can partially automate the evaluation of experiments and the identification of hazardous events from objective measurements.

6 Conclusions

This article presented a dataset and an anomaly-detection framework, for a hardware-in-the-loop (HiL) bicycle simulator. The dataset and processing pipeline were benchmarked on a laptop, a workstation, and an HPC system, demonstrating feasibility on typical research setups, as well as on scalable compute infrastructure.

Future work will implement runtime monitoring, based on the modal logic specification introduced herein; anomaly identification is currently manual, supported by the visual indication of anomalous events in the simulation and by the detected poses. Further topic will include, neural single-camera distance estimation for proximity-based reasoning, neural eye-movement and gaze detection and probabilistic inference (e.g., Bayesian networks or probabilistic neural networks).

The validation will also be extended to more road types and interaction patterns; automated-vehicle effects, which may be subtle in aggregate metrics, will be captured with event-aligned measures (e.g., conflict-related AOIs, time-to-first-fixation, sequential scanning patterns); and the gaze–pose–physiology pipeline will be standardized, with machine learning for data quality control, anomaly detection, and predictive safety indicators.

Overall, the system is a flexible and scalable test environment, for cyclist safety, human–machine interaction and infrastructure design, serving as a research platform, a design decision-support tool and an educational demonstrator.

Acknowledgements

The research was supported by the European Union within the framework of the National Laboratory for Artificial Intelligence (RRF-2.3.1-21-2022-00004).

This research was conducted with the support of the University Research Scholarship Program funded by the Ministry of Culture and Innovation and the National Research, Development and Innovation Fund.

References

[1] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” *arXiv*, 2018.

[2] Y. Tian, Q. Ye, and D. Doermann, “YOLOv12: Attention-Centric Real-Time Object Detectors,” *arXiv preprint arXiv:2502.12524*, 2025.

[3] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” 2020. arXiv:1905.11946.

[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” 2016. arXiv:1506.01497.

[5] W. Abdulla, “Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow,” 2017. GitHub repository.

[6] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, June 2012.

[7] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A Multimodal Dataset for Autonomous Driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] E. R. Corral-Soto, Y. Liu, T. Cao, Y. Ren, and L. Bingbing, “3DArticCyclists: Generating Synthetic Articulated 8D Pose-Controllable Cyclist Data for Computer Vision Applications,” 2025. arXiv:2410.10782.

[9] S. Yaqoob, S. Cafiso, G. Morabito, and G. Pappalardo, “Detection of anomalies in cycling behavior with convolutional neural network and deep learning,” *European Transport Research Review*, vol. 15, p. 9, Mar. 2023.

[10] G. Grigoropoulos, N. Khabibulin, A. Keler, P. Malcolm, and K. Bogenberger, “Detection and Classification of Bicyclist Group Behavior for Automated Vehicle Applications,” pp. 1883–1889, Sept. 2021.

[11] Z. Fang and A. M. López, “Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4773–4783, 2020.

[12] S. Zernetsch, V. Kress, M. Bieshaar, J. Schneegans, G. Reitberger, E. Fuchs, B. Sick, and K. Doll, “Detecting Intentions of Vulnerable Road Users Based on Collective Intelligence as a Basis for Automated Driving,” in *Cooperatively Interacting Vehicles: Methods and Effects of Automated Cooperation in Traffic* (C. Stiller, M. Althoff, C. Burger, B. Deml, L. Eckstein, and F. Flemisch, eds.), pp. 35–87, Cham: Springer International Publishing, 2024.

[13] Y. Belkada, L. Bertoni, R. Caristan, T. Mordan, and A. Alahi, “Do Pedestrians Pay Attention? Eye Contact Detection in the Wild,” *ArXiv*, vol. abs/2112.04212, 2021.

[14] R. Hata, D. Deguchi, T. Hirayama, Y. Kawanishi, and H. Murase, “Detection of distant eye-contact using spatio-temporal pedestrian skeletons,” pp. 2730–2737, Oct. 2022.

[15] H. Murakami, J. Chen, D. Deguchi, T. Hirayama, Y. Kawanishi, and H. Murase, “Pedestrian’s Gaze Object Detection in Traffic Scene,” pp. 333–340, Jan. 2024.

[16] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

[17] K. Khalzaa, S. Karungaru, and K. Terada, "Creation and Testing of Synthetic Datasets for Training Road Scenes Algorithms," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1587–1592, 2023.

[18] K. Khalzaa, S. Karungaru, K. Terada, and T.-A. Chimed-Ochir, "Photo-realistic traffic scene datasets: real and synthetic perspectives for vision-based analysis," *Data in Brief*, vol. 62, p. 111976, 2025.

[19] I. Kotseruba, A. Rasouli, and J. Tsotsos, "Joint Attention in Autonomous Driving (JAAD)," Sept. 2016.

[20] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6261–6270, 2019.

[21] C. Greer, M. Burns, D. Wollman, and E. Griffor, "Cyber-physical systems and internet of things," Tech. Rep. NIST SP 1900-202, National Institute of Standards and Technology, Gaithersburg, MD, 2019.

[22] A. Hammami, A. Borsos, and Á. P. Sándor, "How realistic a bicycle simulator can be? - A validation study," *Multimodal Transportation*, vol. 4, no. 1, p. 100193, 2025.

[23] P. Baranyi and Á. Csapó, "Definition and Synergies of Cognitive Infocommunications," *Acta Polytechnica Hungarica*, vol. 9, pp. 67–83, Jan. 2012.

[24] G. Jocher and J. Qiu, "Ultralytics YOLO11," 2024.

[25] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation," in *Advances in Neural Information Processing Systems*, 2022.

[26] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "DETRs Beat YOLOs on Real-time Object Detection," 2023. arXiv:2304.08069.

[27] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," 2019. arXiv:1906.08172.