OBSERVATION-BASED PEDESTRIAN SCENARIO EXTRACTION FOR VIRTUAL TESTING

Martin Schachner[1], Nadezda Kirillova[2,4], Fabian Weißenbacher[1], Bernd Schneider[1], Horst Possegger[2], Horst Bischof[2], Arno Eichberger[3], Zoltan Ferenc Magosi[3], Jan Dobberstein[5], Thomas Lich[6], Martin Kirchengast[7], Marcus Hennecke[8] and Corina Klug[1]

[1]Vehicle Safety Institute, Graz University of Technology, Austria
[2]Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[3]Institute for Automotive Engineering, Graz University of Technology, Austria
[4]Christian Doppler Laboratory for Semantic 3D Computer Vision, Austria
[5]Mercedes Benz AG, Germany
[6]Robert Bosch GmbH, Germany
[7]AVL List GmbH, Austria
[8]Infineon Technologies Austria AG, Austria

Paper Number 23-0167

Abstract

An overall reduction of pedestrian-vehicle collisions is expected with the market penetration of Advanced Driver Assistant System (ADAS) and autonomous driving (AD) functions. The performance of ADAS is commonly evaluated through virtual scenario-based testing. Hence, scenario catalogs that represent realistic pedestrian-vehicle interactions are needed.

This study shows an approach to automatically extract pedestrian-vehicle scenarios at a selected road intersection, which was observed with a dual-lens stationary observation system. A deep learning-based visual perception pipeline was implemented to reconstruct road user trajectories via state-of-the-art object detection, visual multi-object tracking and object re-identification models. These models were trained and fine-tuned on a manually annotated, diverse dataset, randomly sampled from recordings over multiple weeks. All models were evaluated using common performance metrics. Additionally, localization precision of reconstructed trajectories was assessed using georeferenced ground truth measurements conducted at the intersection. The visual perception pipeline was applied on selected video data and extracted trajectories converted according to the openSCENARIO standard, including a virtual representation of the selected road intersection. The compiled scenarios were further simulated with the openPASS framework.

The results show that pedestrians and vehicles were tracked with high accuracy (Multiple Object Tracking Accuracy $> 83.2\%$) and trajectories were reconstructed with a mean deviation of 0.9 m for pedestrians and vehicle paths with a deviation of 0.68 (SD 0.5) m. The observation system allowed both the obtaining of typical pathways and also speed profiles. An exemplary reconstructed scenario was successfully resimulated in the openPASS framework.

The described approach is promising and can be used to create new scenario catalogs for scenario-based assessments in line with the openSCENARIO standard. Furthermore, the viewpoint of the observation system allows the reconstruction of pedestrian attributes including poses, age or gender, which, alongside an analysis of the recorded pathways and speed profiles with respect to influencing factors, is a focus of ongoing research.

Keywords: camera-based scenario reconstruction, pedestrian behavior, pedestrian-vehicle scenarios, traffic simulation, virtual ADAS evaluation, effectiveness assessment

1 Introduction

Every 5th road user killed in Europe is a pedestrian [11]. To counteract this trend, partner protection by other road users is an issue of great importance. ADAS and AD functions, such as Forward Collision Warning (FWC), Autonomous Emergency Braking (AEB), evasive steering or combinations of these are promising technologies to decrease the number of pedestrian accidents or at least to reduce collision speeds [23, 40], which has a positive effect on pedestrian's injury risks [38]. Scenario-based evaluation is a commonly used method to assess the effectiveness of ADAS for pedestrian safety. The creation of scenario catalogs, consisting of critical pedestrian-vehicle interactions is a major challenge. This effects on the one hand the scenario representation, but also the usage of appropriate data sources. The representation requires a domain-specific scenario description language (SDL), which needs to be in line and interpretable by the underlying simulation environment. A commonly used SDL is openSCENARIO [1], which requires a model of the scenery and the integration of the dynamic entities through storyboards. In order to fill those storyboards, reconstructed real-world accidents can be used according to [8, 16]. In addition to the fact, that accidents are rare events [22], leading to small sample sizes, the pedestrian movement prior to a collision can only be reconstructed with great difficulty and is thus often simplified, *i.e.* assumed to be constant. To compensate this drawback more data on pedestrian behavior and movement is needed.

Camera-based traffic observations are a promising alternative to complement missing information and benefit from new deep learning approaches for automatic scene reconstruction, capable to detect (*e.g.* [20, 34]) and track objects (multiple object tracking (MOT)) over multiple video frames (*e.g.* [48, 51]). Scene reconstruction can therefore be used to better understand the pedestrian behavior in the pre-crash phase [39, 26] but also to derive entire scenario catalogs [50, 4], which incorporate information of the entire scenery, *i.e.* not only of the conflict partners.

There are a variety of different pedestrian observation datasets recorded for specific application purposes. Datasets that record pedestrian movements from a static observation point, *e.g.* [2], mostly serve as benchmarks for tracking algorithms and usually do not provide interaction with other road users, *i.e.* vehicles. Datasets recorded from a vehicle centered view, as shown in [13, 45, 7] have the drawback that trajectories are only recorded over a relatively short time horizon. The datasets published by [50, 4] record pedestrian-vehicle interactions with drones, which make difficult any further determining of pedestrian attributes, such as *i.e.* age [5] or distractions [42, 19], which have an impact on pedestrian behavior. Thus, a trade-off between the level of detail and the overall observability of the scenery must be ensured to provide the necessary details for reconstructing pedestrian-vehicle interactions.

The objective of this study is to present a framework to automatically derive pedestrian-vehicle scenarios capable of integration in common traffic simulation frameworks from a camera-based observation system.

2 Method

For automatic extraction of pedestrian and vehicle trajectories, state-of-the-art computer vision algorithms, consisting of deep learning-based visual MOT and image classification models, are combined to a visual perception pipeline. Different datasets were generated in order to enhance the performance of existing tracking and classification models, but also to evaluate the accuracy of reconstructed trajectories. The traffic observation and the newly generated datasets build the first part of this section, while road network modelling and the visual perception pipeline for trajectory reconstruction build the second part. Details on the simulation of selected reconstructed scenarios complete this section. The overall approach of this and the interplay of the different parts is shown in Figure 1.

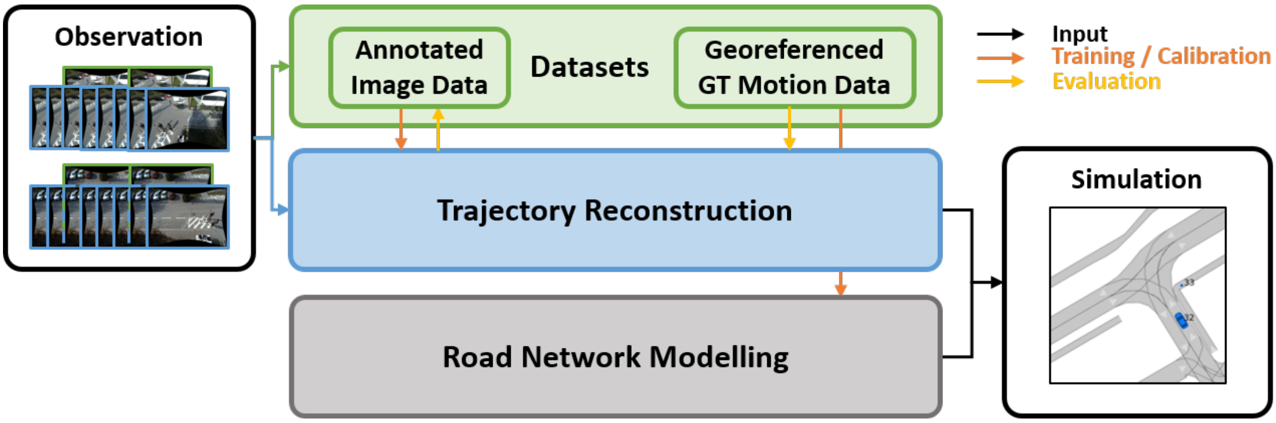

Figure 1. Outline and interplay of different components for automatic scenario extraction with a camera-based traffic observations system. The dataflow among different components is highlighted through arrows. Recorded data from the *observation system* is the input for the *trajectory reconstruction* process (blue frames), selected images (green frames) are combined into *datasets* to enhance (Training/Calibration, orange arrow) and quantify the reconstruction accuracy (Evaluation, yellow arrow). Reconstructed trajectories were aligned with the *road network* (grey) and incorporated in a *simulation* environment (black arrow).

2.1 Camera-based Observation

A robust system was developed for camera-based observation, which enables continuous data extraction even under extreme weather and temperature conditions. As shown in Figure 2, it consists of an Industrial Personal Computer (IPC) inside of a robust control cabinet, on which a dual-lens camera and a long-term evolution (LTE) antenna are mounted. The combined field of view (FOV) of the dual-lens camera thus allows recording of almost 180°, shown in Figure 3.

The camera-based observation system was installed at an intersection of two private service roads, forming a T-junction, at the campus Inffeldgasse of Graz University of Technology. At the observation point, the traffic operates in accordance with the Austrian road traffic regulations, which implies right-hand driving and giving priority to the right at intersections. A speed limit of max. 20 km/h applies within the area. Furthermore, ground markings and road signs give information about towing zones as well as other prohibitions or bans. Sight obstructions, caused by parked vehicles as well as a large-scale art installation (a frame structure spanning the road), lead to a potential threat for pedestrians and potentially to interesting and frequent interactions with vehicles.

In order to record the intersection appropriately with the camera-based observation system, it was mounted at the frame structure of the art installation at a height of approximately 5 m, as shown in Figure 2. The cameras were aligned accordingly to best observe the events at the road intersection, which is shown in Figure 3.

Since the projection of 3D real world information onto a 2D image discards metric information, the dual-lens camera must be properly calibrated to allow recovering of 3D units from image-based measurements. To this end, the system was calibrated intrinsically (using the calibration target and framework from [12], which extends [6]) which allows image rectification, *i.e.* to correct the inaccuracies caused by the inherent distortion of the optical lenses (which is most notable at the image border). To align the rectified camera images with a common World Coordinate System (WCS) denoted as O_W, *i.e.* extrinsic calibration, the AprilTag [36] framework was used. These calibration markers can be robustly detected outdoors and yield a sufficiently low pose estimation error. In particular, the translation and rotation errors over 20 consecutive video frames are $\leq 1.5\,\mathrm{mm}$ and $\leq 0.1°$, respectively. In this study the WCS was calibrated in such a manner that an XY-plane is placed on the ground (*i.e.* at $Z = 0$), and the Z-axis is pointing up, see Figure 3.

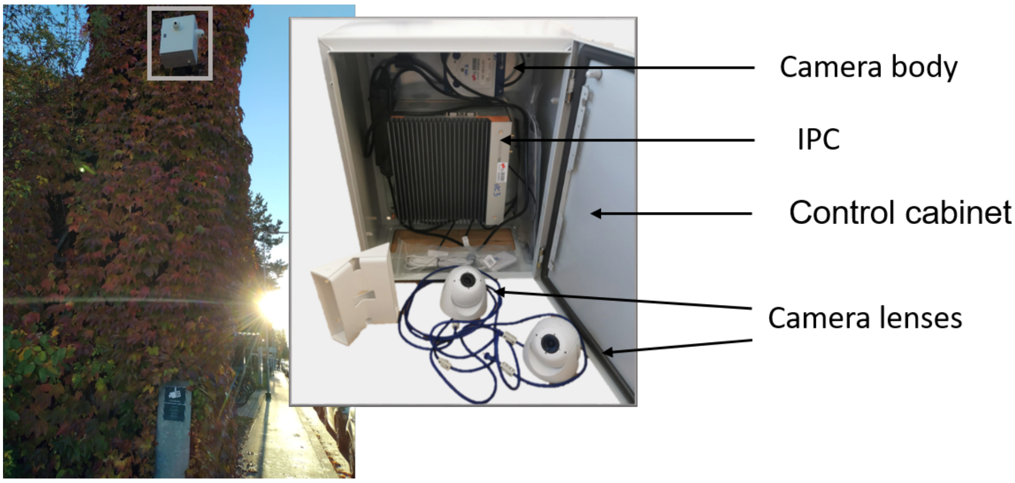

Figure 2. Mounted observation system at campus Inffeldgasse. The left image shows the observation system mounted at the art installation at the selected intersection. A sign informs pedestrians about the project and the legal basis of data collection. The right image shows the main components of the observation system, which consists of an IPC, connected to a Mobotix S16B camera body and the two camera lenses. An LTE antenna is mounted on the side of the control cabinet.

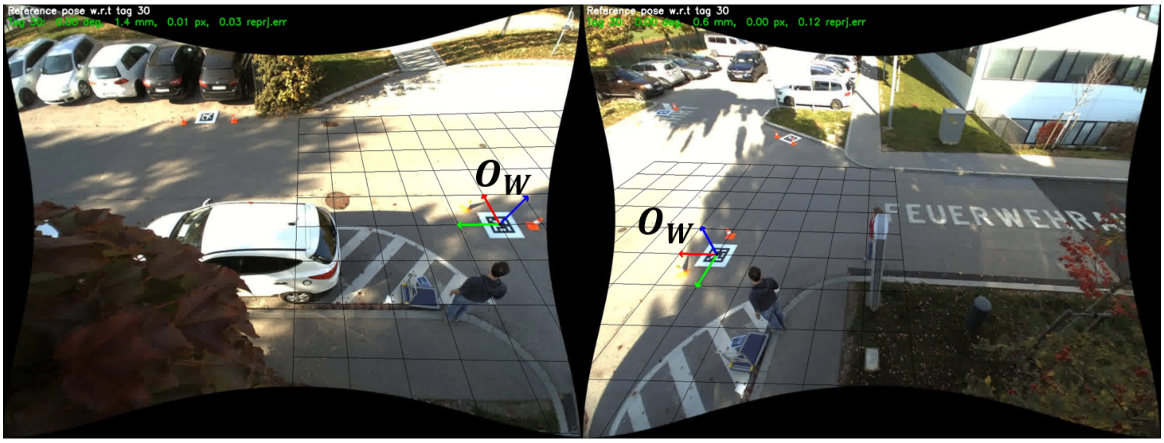
Reference pose w.r.t tag 30
Tag 30: 0.00 deg, 1.4 mm, 0.01 px, 0.03 reprj.err
Reference pose w.r.t tag 30
Tag 30: 0.00 deg, 0.6 mm, 0.00 px, 0.12 reprj.err
O_W
O_W
FEUERWEHR

Figure 3. Left and right FOV of the installed observation system at the selected intersection. The extrinsic camera calibration defines the WCS O_W, which is illustrated via arrows pointing along the major axes (red: X-axis, green: Y-axis, blue: Z-axis).

2.2 Datasets

The camera placement of the observation system allows recording of the events at the selected intersection, but it does not resemble the viewing angle of publicly available datasets. Thus, the object detection, tracking and re-identification models had to be properly adjusted via finetuning. For this, we collected a sufficiently diverse image dataset. In addition to the collected image data, dedicated experiments were performed, in which georeferenced locations of vehicles and pedestrians were measured. This data was used to evaluate the accuracy of the trajectory reconstruction process.

2.2.1 Image Data

In order to generate a suitable dataset for fine-tuning and performance evaluation of a MOT algorithm, a dataset consisting of 10,800 frames was created from the recorded videos at the observation point. Each frame was manually annotated using the CVAT tool [43], where each annotation consists of an object's bounding box, as well as other relevant attributes, such as age category, gender, personal mobility device (scooter, bicycle, *etc.*) and potential distractions caused by smartphones or headphones. One part of the dataset, consisting of 7,200 frames, acts as training, the other as evaluation data. Due to the demographic conditions at the selected intersection, some attributes and objects are underrepresented, when sampling uniformly from the sample images (*e.g.* children, adolescents). To cope with this issue, particular interesting samples have been searched manually.

2.2.2 Ground Truth Motion Data

For estimating the accuracy of the trajectory reconstruction process, geolocations of pedestrians and vehicles were measured within dedicated tests. The geolocations were measured in both cases with an inertial navigation system (INS), using the Global Positioning System (GPS). Going forward, this measured data will be denoted as ground truth (GT) motion data. The temporal synchronization between video and the GT motion data was established through timestamps, associated to each video frame and the measurement, respectively.

Pedestrian The measurements of pedestrian GT motion data were designed to further allow for estimating the accuracy of the camera projection with respect to the WCS. For measuring the GT motion data, the INS (type: OxTS RT3000 v2 [30]) was mounted on a trolley. In order to be able to identify the INS in the recorded video frames using the AprilTag [36] framework, the trolley was equipped with a calibration tag, offset relative to the INS (in the reference frame of the measurement trolley). The INS base station was placed near the measurement area, with a sufficient distance to surrounding buildings that could have shielded the GPS signal and thereby could have introduced additional uncertainties. The developed measurement setup is shown in Figure 4.

In the experiment, the trolley was moved in the scene in such a way that it was visible in the video recordings. By this means the area of interest was covered by a grid of measurements with a spacing of about 1 m, resulting in around 200 distinct georeferenced measurement positions. Each position was captured for approximately 10 s, during which the trolley remained stationary. This made it easier to extract the particular image point and the corresponding GT position data. For the further usage, the position data of each measurement position was time-averaged to reduce measurement noise.

Vehicle For recording GT vehicle motion data, a dedicated test vehicle has been adapted to the needs of this investigation. Figure 5 sketches the vehicle setup schematically. Its setup consists of

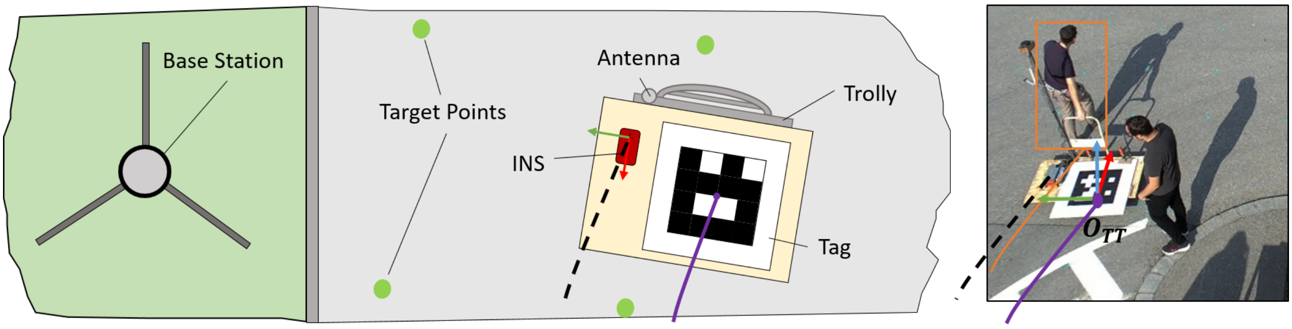

Figure 4. Schematic representation of the measurement setup installed on the movable trolley. The trolley was equipped with a calibration tag, the INS and its GPS antenna. Illustrated are the measured GPS track (black dashed), as well as the projected tag position (purple) and the trajectory resulting from the tracked operator (orange).

a data acquisition unit (DAU) (type: dspace AUTERA autobox operated by Intempora RT-MAPS 4 data acquisition software) which collects and synchronizes data from different sensor sources. A combination of accurate GPS-RTK (type: Novatel OEM 6, corrected by APOS service) and intertial measurement unit (IMU) (type: Genesys ADMA-III) is used to record trajectories and dynamic driving states (position angle, speeds, acceleration *etc.*). This is complemented by a Light Detection and Ranging (LiDAR) sensor (type: OUSTER OS1-128) which operates with a resolution of 128 x 1024 at 10 Hz and triggers the recording of the installed camera system (type: IDS - UI-5240CP with TAMRON M118FM08 lens).

Within the experiment the test vehicle was driven through the intersection several times, covering all dedicated routes given by the road network, as shown in Figure 6. The test drives were made in the speed range foreseen for the intersection. Further, recordings include typical interactions with other road users, especially pedestrians, in which the driver gives right of way and vice versa.

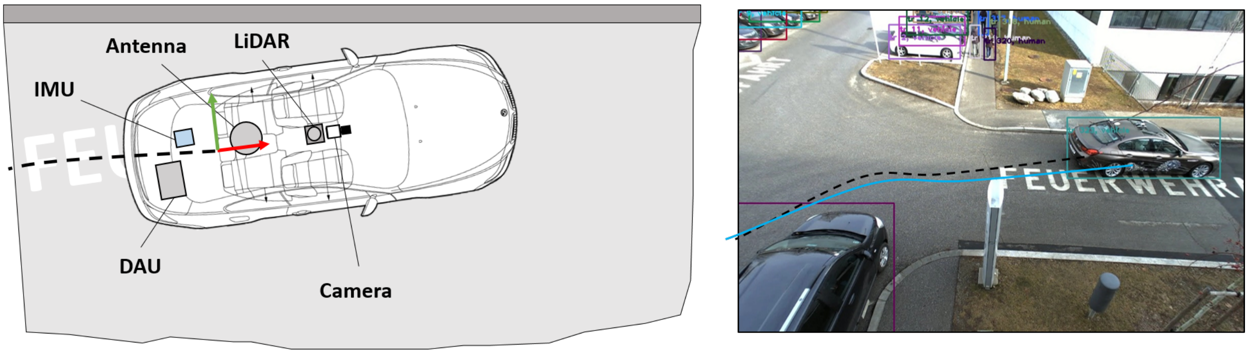

Figure 5. Schematic representation of the measurement setup installed in the test vehicle. The vehicle was equipped with a DAU, an IMU for measuring geolocation as well as a LiDAR sensor and a camera system. Shown are the GPS track (black dashed) that was measured in the conducted test and the projected trajectory resulting from the scenario reconstruction (blue).

2.3 *Road Network Modelling*

The static environment of the selected observation point (discussed in Section 2.1) serves as a basis for the agent simulations and have been digitized such that they are in line with the specifications of the openSCENARIO [1] standard. In the following the 3D scene representation is explained as well as its alignment with the reconstructed trajectories in WCS.

2.3.1 3D Scene Representation

For the modelling, freely available, high quality geographic data of the selected intersection was used. This includes the direct proximity to the mounted observation system within the observation (*e.g.* intersection, crosswalks, etc.) as well as the adjacent road network within a radius of about 100 m. Orthophotos, retrieved from [15], as well as surface and terrain information, retrieved from [14], were used as data sources. As a pre-processing step, both data sources were merged and processed with QGIS [32]. For the creation of the 3D scenes, RoadRunner [29] was used. The construction of the road network, including sidewalks, was done manually, based on orthophotos of the observation point. The resulting 2D road network was supplied with terrain information and exported as an openDRIVE file. A visualization of the resulting 3D openDRIVE road network is shown in Figure 6.

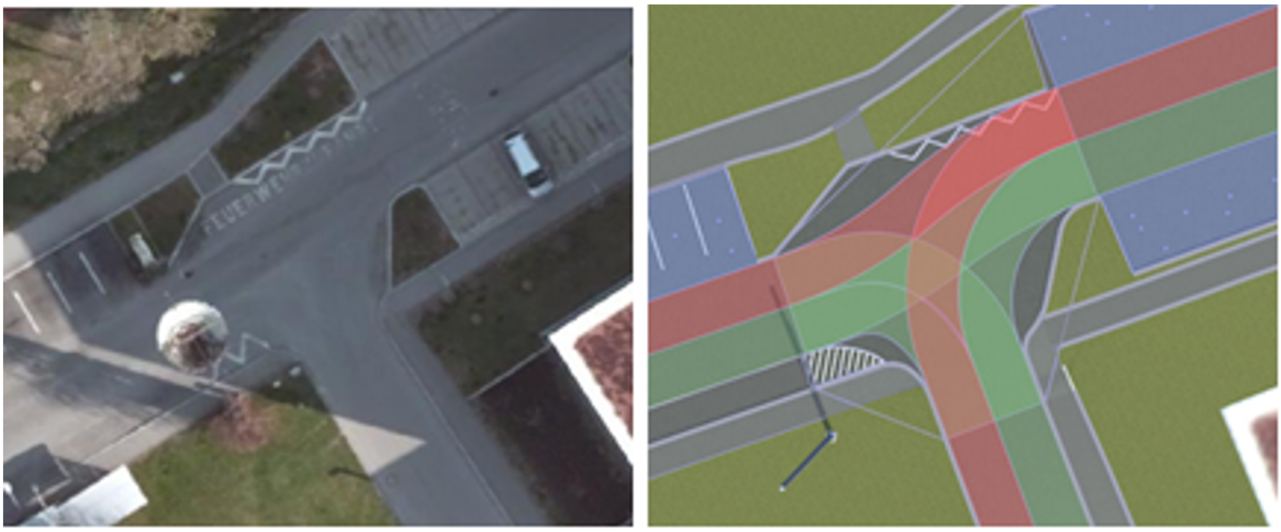

Figure 6. Modelling of the Inffeldgasse observation site. In the left image the orthophoto of the intersection area is shown. In the right image a 3D model of the observation point including the overlay of the openDRIVE network and the driveable routes.

2.3.2 Coordinate System Alignment

As described in Section 2.1, the WCS O_W is determined by the extrinsic camera calibration, which is used to recover 3D information of the road users, *i.e.* projecting trajectories to the ground plane. Reconstructed trajectories need to be transformed from the WCS to the OpenDRIVE Coordinate System (OCS) O_{OD} in order to obtain the representation of the trajectory in openSCENARIO. This OCS is bound to the georeferenced openDRIVE representation of the scene and therefore its coordinate axes are aligned to the cardinal directions following the east-north-up (ENU) convention.

GT motion data from the trolley measurements was used to determine the transformation T_{W2OD} between WCS and OCS. Therefore a third, intermediary coordinate system (ICS) O_I was defined by a reference tag at a selected measurement location I. Through the georeferenced position and orientation of I, the transformation of O_I relative to O_{OD} is known. Further, the transformation from O_I to O_W can be computed using the AprilTag framework. T_{W2OD} is thus defined by chaining these transformations. It should be noted that only the Z-rotation angle (heading) of I was used, since pitch and roll angle were both relatively small, and their influence on the X/Y position was thus deemed negligible (< 5 cm).

2.4 Trajectory Reconstruction

The implemented visual perception pipeline for automated trajectory reconstruction consists of four building blocks, as shown in Figure 7. The central part includes the MOT algorithm, which localizes the road users, *i.e. pedestrians* and *vehicles*, throughout video frames while maintaining their identitys (IDs). Tracked road users in image coordinates then have to be mapped to the scenery representation (ground plane (WCS) and further to OCS), which forms the second part of the pipeline. To obtain consistent trajectory IDs across the partially overlapping FOVs of the dual-lens camera, the third step is to match corresponding road user trajectories across both FOVs. Finally, a post-processing step suppresses potential measurement noise from the reconstructed trajectories.

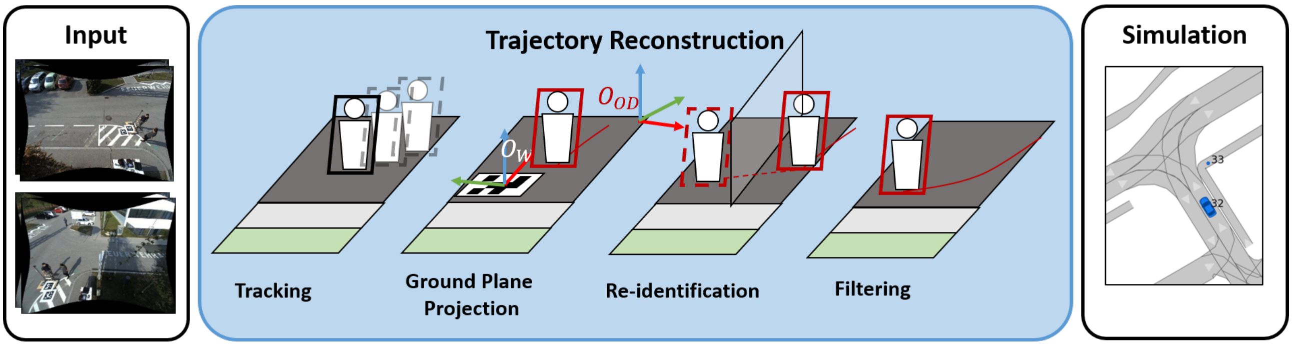

Figure 7. Schematic representation of the visual perception pipeline for trajectory reconstruction. The input frames from the cameras' FOVs are processed by the four building blocks, *multiple object tracking*, *ground plane projection*, *re-identification* and *trajectory filtering*.

2.4.1 Multiple Object Tracking-by-Detection

The MOT algorithm used follows *the tracking-by-detection paradigm* and provides a favorable balance between high accuracy and low run-time. It leverages a state-of-the-art single-stage object detector, *i.e.* YOLOv5 [20, 34], which generates object hypotheses in the form of bounding boxes (*i.e.* rectangular regions which likely contain an object) per frame. These object hypotheses are then temporally linked to object trajectories via an implemented multi-class capable extension of the DeepSORT [48] MOT algorithm. For improved robustness, the standard appearance feature estimation in DeepSORT was replaced by the re-identification model OSNet [51], which performs favorably for varying object sizes and is thus more suitable for deployment in this study's observation scenario.

The output of the tracking step is a sequence of bounding boxes for each road user k over time t, *i.e.*

$$\mathbf{O}^{(k)} = \left(O^{(k)}_{t_{init}}, \dots, O^{(k)}_{t_{end}} \right), \tag{1}$$

where a bounding box $O^{(k)}_t = (x^{(k)}_t, y^{(k)}_t, w^{(k)}_t, h^{(k)}_t)$ is defined by its top-left corner coordinates, width and height, respectively. All units are in pixels. The set of all estimated trajectories is denoted as

$$\mathcal{O} = \{\mathbf{O}^{(k)}\}_{k \in [1,\dots,N]}, \tag{2}$$

where N is the number of all detected road users.

2.4.2 Ground Plane Coordinates from Image-based Measurements

The image-based bounding box trajectories $\mathcal{O}$ need to be projected into the WCS to obtain 2D road user trajectories. For the projection, the widespread *central perspective projection* model (*pinhole camera*) is assumed, which follows the collinearity principle, *i.e.* each real-world point is projected along a straight line through the projection center (the camera's optical center) onto the image plane [18]. Using both intrinsic and extrinsic calibration of the camera, a homography (projective collineation) can be derived which allows mapping image coordinates onto a reference plane in the world coordinate system. Since the object foot points can be easily estimated from the image-based detection results, the world ground plane was chosen at $z = 0$ as reference plane for this projection.

In particular, given an object's bounding box $O_t^{(k)}$, we leverage the scene geometry given by the extrinsic calibration to compute the object's *orientation vector*. This allows accurate location of the *foot* (bottom) and *head* (top) points of the road user by intersecting the orientation vector with the edges of the corresponding bounding box, as illustrated in the left and middle image of Figure 8, the right image shows exemplary trajectories superimposed on a bird's eye view image. To obtain the corresponding location in OCS $\mathbf{x}_t^{(k)} = (x_t^{(k)}, y_t^{(k)}, 0)$ (measured in mm), the derived *foot* points are projected onto the world ground plane, *i.e.* WCS representation and further transformed to the OCS, as described in 2.3.2. Thus, the reconstructed trajectory signal $\mathbf{x}^{(k)}$ of a road user k is

$$\mathbf{x}^{(k)} = \left(\mathbf{x}_{t_{init}}^{(k)}, \ldots, \mathbf{x}_{t_{end}}^{(k)}\right). \tag{3}$$

In order to represent the road user movements within openSCENARIO, the reconstructed trajectory $\mathbf{x}^{(k)}$ is further filtered as described in 2.4.4,

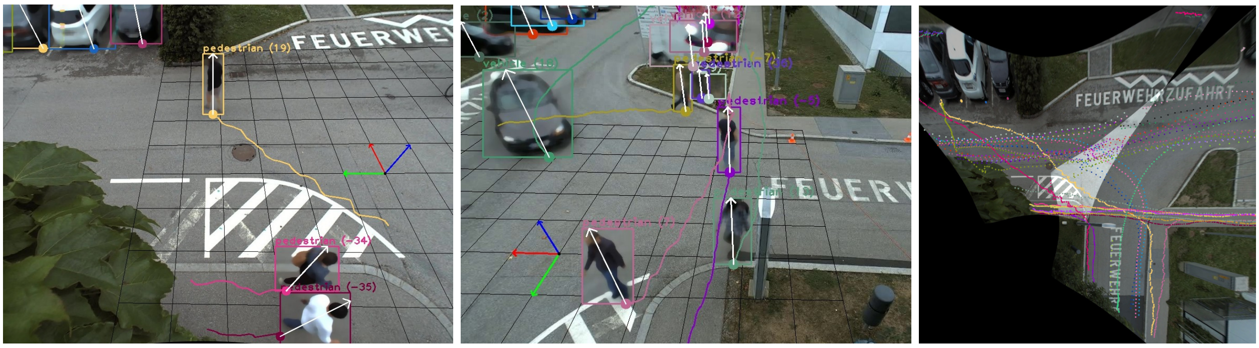

Figure 8. Multi-object multi-class tracking visualization with object *orientation vector* estimation and road user's *foot point* derivation. The colors of the bounding boxes and trajectories correspond to the object's instance ID. The right image shows exemplary trajectories after projecting image-based localization results onto the world's ground plane. The bird's eye view image was obtained by projecting the camera image pair onto the same plane.

2.4.3 View-consistent Trajectory IDs via Re-identification

The observation system uses two partially overlapping FOVs to cover a larger area of the road intersection. To obtain the most accurate localization results, tracking-by-detection is performed in each FOV independently. For consistent trajectory IDs throughout the *two-stream* scene, a road user re-identification approach is applied to establish correspondences between them in different FOVs. The method reuses the appearance features already extracted by the OSNet re-identification model from the tracking step (as discussed in Section 2.4.1) which allows saving computational resources. For each road user, a feature gallery for 60 video frames, which is 5 seconds by recording at 12 fps, is kept and used in the following feature matching step. To predict an object transition between the two FOVs, temporal and spatial information is used in the same coordinate system, *i.e.* per frame coordinates projected onto the world ground plane (see Section 2.4.2). Thus, when a new road user is detected in the current FOV, and its real-world trajectory lies in the area of the second FOV, the feature galleries of these objects are matched across both FOVs and assigned a consistent trajectory ID upon finding correspondences in the galleries.

2.4.4 Velocity Estimation through Trajectory Filtering

The output of the previous steps consists of time-dependent trajectory signal $\mathbf{x}^{(k)}$ in the OCS reference frame (see Equation (3)). Due to the nature of the tracking algorithm (see Section 2.4.1 and 2.4.2), velocity information is not provided explicitly, and the position estimates are related to the bounding box of the objects, which introduces additional measurement noise. In order to retrieve meaningful trajectories and velocity profiles of the tracked road users, a constant acceleration Kalman filter [21] followed by a Rauch-Tung-Striebel (RTS) smoother [33] is applied on the trajectory signal $\mathbf{x}^{(k)}$. This approach benefits from improving $\mathbf{x}^{(k)}$, while yielding an estimate for the tracked object's velocity $\dot{\mathbf{x}}^{(k)}$ and acceleration $\ddot{\mathbf{x}}^{(k)}$. The chosen Kalman filter models the X and Y-component of the object's motion independently using position, velocity and acceleration as state variables. The acceleration may change between time steps, based on a discrete white noise model, where the amount of change is controlled by the estimated maximum jerk ($\dot{a}$) of the tracked object. The attached RTS smoother exploits the fact that $\mathbf{x}^{(k)}$ is already known for the whole trajectory when running the filter operation.

When setting up the filter, it was assumed that the measurement noise and the covariance of the position state variable can be estimated with the position accuracy of the tracking algorithm, Δx. Further it has been assumed that the maximum absolute values for velocity and acceleration, v_{max} and a_{max} respectively, can be used to estimate the covariances of the corresponding state variables as $\sigma_v^2 \approx (v_{max}/3)^2$ and $\sigma_a^2 \approx (a_{max}/3)^2$ [24].

2.5 Reconstruction Accuracy Estimation

For the accuracy evaluation of the trajectory reconstruction process, the georeferenced GT trajectories $\hat{\mathbf{x}}^{(k)}$ were compared with the reconstructed $\mathbf{x}^{(k)}$. As a pre-processing step, the measured data had to be aligned and time-synchronized for the setups, *i.e.* the reconstructed trajectories, as described in Section 2.4, had to be linearly interpolated, resulting in a sampling frequency of 100 Hz. Since each $\hat{\mathbf{x}}^{(k)}$ is accompanied by its timestamps, they were used to extract the corresponding frames from the video material. For time ranges in which the GT motion has been recorded in the vicinity of the selected intersection, the video recordings where searched for a tracked object $O_t^{(k)}$ (vehicle, pedestrian), which corresponds to the georeferenced GT motion data. To extract $\mathbf{x}^{(k)}$, the trajectory reconstruction process as described in Section 2.4 has been applied. At this point, it should be mentioned, that for the recordings with the measurement trolley (see Section 2.2.2), two different signals have been retrieved. Sequences, in which the trolley and operator moved in unison offered the opportunity to test the reconstruction process in the wild, *i.e.* on a pedestrian $\mathbf{x}^{(k)}$ as compared to an idealized object $\tilde{\mathbf{x}}^{(k)}$, *i.e.* the center of the tracking tag. This setup is particularly helpful for investigating how bounding box related effects affected the tracked trajectory of the pedestrian, since the tag is tracked with high accuracy and its image position is clearly defined.

Additionally, $\tilde{x}^{(k)}$ was used to test and tune the performance of the Kalman Filter that was subsequently used to smooth reconstructed trajectories and provided velocity estimates, as discussed in Section 2.4.4.

2.6 Pedestrian-Vehicle Scenario Simulation using openPASS

The specifics are described in the following for simulating the observed scenarios in the simulation platform openPASS [47, 9]. Originally, the term openPASS formed a backronym for "Open Platform for the Assessment of Safety Systems". It has since been expanded beyond the scope of safety alone, and extends to the assessment of any kind of ADAS and AD function, using the standards OpenDRIVE, OpenSCENRIO and OpenSimulationInterface. In line with this, the simulation aims to provide vehicle and pedestrian agents that are modelled through system definitions and follow model-based design approaches. The agents are therefore composed by motion dynamics models and may consist of interconnected subsystems, such as ADAS functions. openPASS requires a certain information flow and thus files necessary to perform a simulation study. Besides meta information for the simulation (provided in *SimulationConfig*) and the agent models (*ProfilesCatalog*, *SystemConfigs*), which have to be passed, peculiarities of the scenario description and the properties of the pedestrian and the vehicle agent should be summarized in short form.

Scenario Description using OpenSCENARIO The scenario file describes the traffic situation following the OpenSCENARIO standard [1]. The scenery used was modelled in Section 2.3 and was incorporated into the scenario via its corresponding openDRIVE description. The storyboard, indicating the interplay of road users and the temporal development of the scene has been modelled in accordance to the openPASS PreCrash Matrix (PCM) use case. The behavior is therefore composed as a *FollowTrajectoryAction*, which consists of the trajectory description, sampled with 100 Hz. The same method of trajectory interpolation as described in 2.4 has been used. Furthermore, the scenario file includes a link to a *ProfilesCatalog*, which describes the underlying algorithmic models of the spawned road users.

Pedestrian Agent The pedestrian agent is relatively simple. Prescribed trajectories that are continuous in their velocities and accelerations are defined as a *FollowTrajectoryAction* in the Scenario file. The internal implementation of a trajectory follower forces the pedestrian agent to move to the defined position in time.

Vehicle Agent The vehicle agent is similar to the model used in [40]. The target trajectory is described by the *FollowTrajectoryAction*, defined for each simulated agent in the storyboard of the scenario file. This given target trajectory is passed to a *route control algorithm*, which controls brake, throttle and steering signals based on the deviation between the actual and target state. This behavioral model affects the dynamics of the two-track model and the dynamics of the chassis, which simulate the suspension of the vehicle due to inertia forces.

3 Results

The first the results of the described methodology are presented in the following. The main focus is on the accuracy of the trajectory reconstruction process. Since this visual perception pipeline is significantly influenced by the MOT, its performance has been assessed separately. Further, the reconstructed trajectories have been assessed as described in Section 2.5. This section is complemented with first simulation results, which were generated by using the derived scenario in openPASS.

3.1 Multiple Object Tracking-by-Detection

Since MOT is a complex task consisting of detection, localization and association, multiple metrics should be considered to evaluate the performance of a multi-class multi-object tracking-by-detection approach. We employ the widely adopted multiple object tracking accuracy (MOTA) [3], identification F_1 score (IDF_1) [37] and trajectory quality [25] measures. Since it has recently been shown by [28] that these measures are often biased towards specific components of a tracking system (*e.g.* CLEAR-MOT focuses on detection and localization, while IDF1 focuses on association), we additionally report the results in terms of higher order tracking accuracy (HOTA) [28], which explicitly addresses these biases of existing measures.

The results of the quantitative multi-class multi-object tracking-by-detection evaluation are shown in Table 1. The tracking performance for both pedestrian and vehicle classes achieves high MOTA rates, *i.e.* 83.24% and 92.50%, respectively. Moreover, for the pedestrian class, which is the most important class due to their high vulnerability, the IDF_1 and HOTA scores are also at state-of-the-art levels, *i.e.* 89.39% and 70.61%, respectively. The lower IDF_1 score for the vehicle class can be mostly attributed to identity switches at parking area (visible at the top border of the right FOV). There, vehicle detections are significantly more unstable due to the high degree of occlusions, *i.e.* both the traffic sign, as well as other parked vehicles occlude distant cars which leads to frequent detection failures. Consequently, this causes identity switches for the occluded cars. As this only affects the cars parked at the far end of the camera's FOV, it does not impede the performance of our system to extract trajectories of moving and interacting road users.

Object class	**GT**$_{\mathrm{det}}$	**TP**$^{\uparrow}_{\mathrm{det}}$	**FN**$^{\downarrow}_{\mathrm{det}}$	**FP**$^{\downarrow}_{\mathrm{det}}$	**IDF1**$^{\uparrow}$	**MOTA**$^{\uparrow}$	**IDsw**$^{\downarrow}$	**GT**$_{\mathrm{traj}}$	**MT**$^{\uparrow}_{\mathrm{traj}}$	**PT**$^{\uparrow}_{\mathrm{traj}}$	**ML**$^{\downarrow}_{\mathrm{traj}}$	**HOTA**$^{\uparrow}$
Pedestrian	3634	3180	454	156	89.39%	83.24%	4	41	26	13	2	70.61%
Vehicle	30496	30098	398	1887	44.46%	92.50%	3	88	87	1	0	56.85%

Table 1. Quantitative results of the multiple class multiple object tracking-by-detection method on our image data. $\mathrm{GT}_{\mathrm{det}}$, $\mathrm{TP}_{\mathrm{det}}$, $\mathrm{FN}_{\mathrm{det}}$ and $\mathrm{FP}_{\mathrm{det}}$ - the numbers of ground truth, true positive, false negative and false positive detections, respectively; IDF_1 - tracking Identification or association accuracy score; MOTA - multiple object tracking accuracy; IDsw - the number of trajectory ID switches; $\mathrm{GT}_{\mathrm{traj}}$, $\mathrm{MT}_{\mathrm{traj}}$, $\mathrm{PT}_{\mathrm{traj}}$ and $\mathrm{ML}_{\mathrm{traj}}$ - the numbers of ground truth, mostly tracked, partly tracked and mostly lost trajectories, respectively; HOTA - higher order tracking accuracy. $^{\uparrow}$ and $^{\downarrow}$ denote that higher/lower values correspond to better performance.

3.2 Reconstruction Accuracy

In general, it is hard to evaluate the similarity of two trajectories. One of the most commonly used approaches is the calculation of spatial distances between temporal corresponding points, also referred as lock-step Euclidean distance (LSED) Eu, as well as the calculation of the dynamic time warping (DTW), which is well suited to compare paths [46]. Since the trajectories used for the accuracy evaluation have different lengths, the mean μ_{Eu} and μ_{dtw} as well as the standard deviation σ_{Eu} and σ_{dtw} over all considered point pairs have been used for comparison. Besides a path reconstruction, it is essential to obtain accurate speed profiles of the road users. For the comparison of the speed profiles resulting from the trajectories, the cross-correlation $\rho^{(k)}_{\hat{v},v}$ between $\hat{\mathbf{v}}^{(k)}$ and $\mathbf{v}^{(k)}$ was used and an mean correlation μ_ρ calculated.

The deviation by means of the DTW, are promising and circumvent those effects by comparing paths.

3.2.1 Optimum

The tracked tag positions of all measurement positions (see Section 2.2.2) were mapped into the OCS separately for each FOV using the approach presented in Section 2.3.2 and compared against the recorded GPS measurements (also mapped into the OCS). The resulting deviations in the XY-plane of the OCS are plotted in Figures 9 and 10. For the left FOV, we were able to reconstruct the tracked tag positions with a mean absolute error (MAE) of 0.2 m, a median deviation of 0.14 m and a maximum absolute error of 0.63 m.

For the right FOV, we achieved a MAE of 0.17 m and a median deviation of 0.1 m. The maximum absolute error was higher with 0.95 m, but this can generally be attributed to outliers, since for approximately 95 % of all measured positions, the absolute deviation is actually $\leq$ 0.3 m. These outliers correspond to measurement positions that lay far from the camera and close to the edge of the image (as shown in the right pane of Figure 10) where the lens distortion is highest, thus making them especially susceptible for re-projection errors.

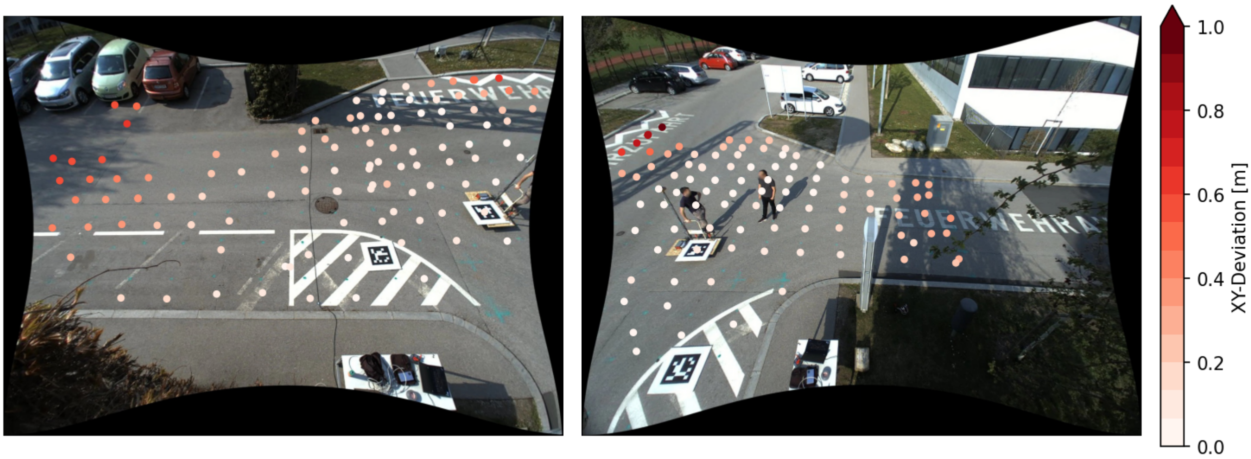

Figure 9. Reconstruction accuracy of the tag position overlaid on camera views. The left and right images show the deviations between GPS measurements and the tracked tag position (transformed via T_{W2OD}) for the left and right FOV respectively.

3.2.2 Pedestrian

Reconstructed trajectories were compared against the recorded ground-truth trajectories for 25 shared movement sequences of the measurement tag and trolley operator (pedestrian). It was found that the measurement tag position was reconstructed with a mean LSED of 0.2 m, which was in line with the optimal reconstruction accuracy of 0.17 m to 0.20 m that was established using static targets (see Section 3.2.1). The positional accuracy of the operator trajectories, in contrast, was significantly lower with an average LSED of 0.9 m and with maximum absolute errors reaching as high as 2.5 m compared to 1.0 m for the tag trajectories. The poorer performance for the operator trajectory can be explained to a great extent, as the effect of the distance between the operator and the measurement tag, which introduces a systematic error that could not be corrected during the analysis. For the

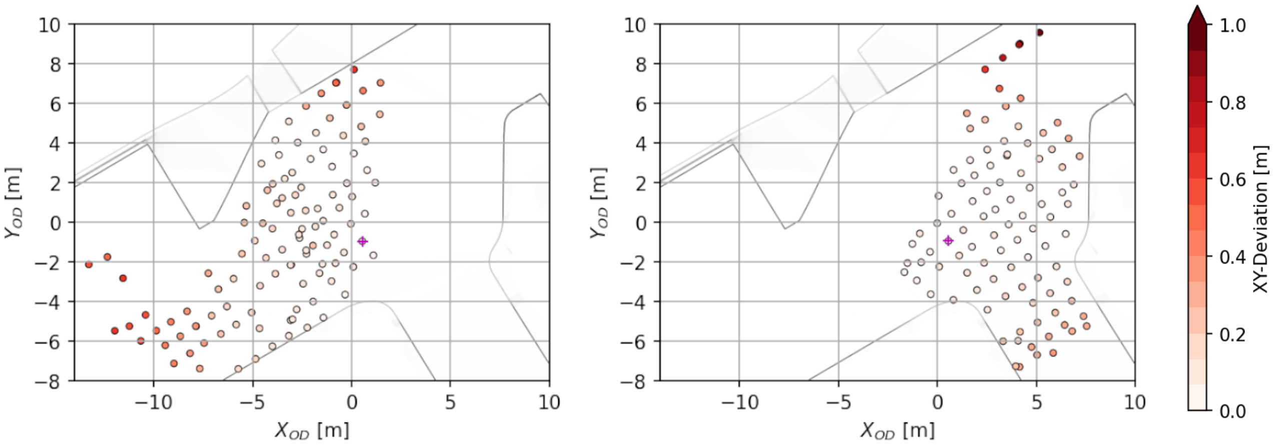

Figure 10. Reconstruction accuracy of tag position overlaid on road borders. The left and the right images show the deviations between GPS measurements and the tracked tag position (transformed via T_{W2OD}) for left and right FOV respectively. The position of the intermediary coordinate system is marked with a magenta cross.

comparison of velocities, the offset between operator and tag should have had less of an impact, assuming that it stayed approximately constant. On examining the calculated velocity accuracies, this has been the case: the LSED of the operator velocities was higher than that of the tag velocities, 0.27 m/s for operator versus 0.07 m/s for the tag, but the relative difference between these values was smaller than between the corresponding positional LSEDs. This can also be seen in Table 2, which summarizes the position and velocity accuracy statistics.

Figure 11 gives a qualitative comparison between tracked and reconstructed trajectories of reference tag and trolley operator for a selection of analyzed movement sequences. It can be seen that the recorded operator trajectory was significantly noisier than the tag trajectory, underscoring the need for Kalman filtering when confronted with realistic data. The Kalman filter was designed to filter out the high-frequency noise in the pedestrian tracking data, which is most apparent in the trajectories for sequences 10, 12, 23 and 24. This noise is generated by oscillations in the bounding box, and is most likely caused by foot movement and the pedestrian's legs being partially occluded by the measurement trolley. As can be seen from the right side of Figure 11, the filter generally succeeded in removing this type of noise.

	Position		Velocity		
Data source	μ_{Eu} $[m]$	$Eu_{\max}$ $[m]$	μ_{Eu} $[m/s]$	$Eu_{\max}$ $[m/s]$	μ_{ρ}
Optimal (tag)	0.2	1.0	0.07	0.7	0.97
in the wild (pedestrian)	0.9	2.5	0.27	2.2	0.77

Table 2. Quantitative results of the pedestrian trajectory reconstruction accuracy. μ_{Eu} - mean LSED, $Eu_{\max}$ - maxmium LSED, μ_{ρ} - mean velocity profile correlation

3.2.3 Vehicle

A total of 18 reconstructed trajectories were compared with their georeferenced GT measurement. The average LSED showed a reconstruction accuracy of 1.391 (± 0.77) m and the DTW 0.675 (± 0.51) m. Vehicle speed could be reconstructed with a MAE of 0.34 (± 0.38) m/s. Reconstructed and

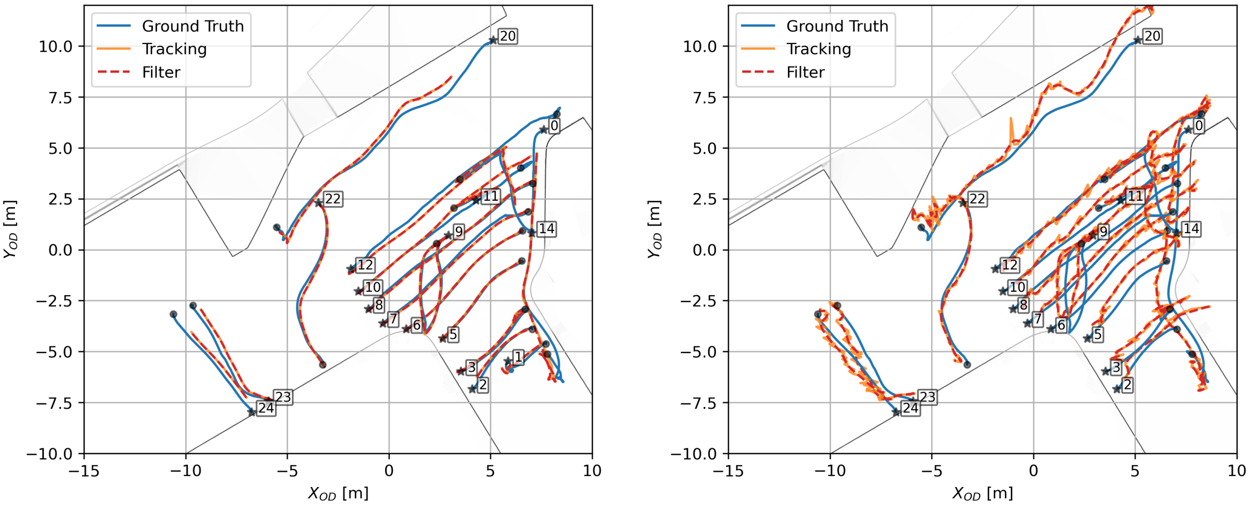

Figure 11. Qualitative comparison between reconstructed trajectories (red dashed lines) and geo-referenced ground-truth data (blue) for selected movement intervals. Left: trajectories based on the measurement tag. Right: trajectories of the person that operated the measurement trolley. The beginning and end of all sequences is marked by star and circle symbols, respectively. Best viewed on screen.

measured velocities showed a correlation of about 0.94 on average. The high deviation by means of LSED can be reasoned by the determination of the vehicle's representative point, which is a function of the bounding box. Therefore a systematic error in the distance between the GPS measured point and the reconstructed point effects the result. The deviation by means of DTW are however promising and circumvent those effects by evaluating path similarity. Furthermore, these results have been divided by means of the six different paths, which could possibly be taken by a vehicle on the basis of the road layout. The obtained results per path are provided in Table 3, selected trajectories per path are shown in Figure 12.

Scenario	**Nr**	**Nr Points**	μ_{Eu} $[m]$	$\sigma_{Eu}[m]$	μ_{dtw} $[m]$	$\sigma_{dtw}[m]$	$Eu_{\max}[m]$	μ_{ρ}
1-2	5	4944	1.406	0.518	0.54	0.414	2.425	0.974
1-3	4	4805	1.324	1.013	0.894	0.521	3.203	0.945
2-1	4	3335	1.079	0.475	0.383	0.305	2.378	0.884
2-3	2	1080	1.252	0.309	0.57	0.379	1.543	0.939
3-1	3	3040	1.911	0.799	0.841	0.627	3.289	0.941
3-2	3	1265	1.284	0.688	0.837	0.463	2.792	0.967
all	21	18469	1.391	0.767	0.675	0.511	3.289	0.942

Table 3. Quantitative results of the vehicle trajectory reconstruction accuracy. Results have been subdivided into different scenarios, based on the paths. μ_{Eu} - Mean LSED, σ_{Eu} - standard deviation of LSED, μ_{dtw} - Mean DTW, σ_{dtw} - standard deviation of DTW $DTW_{\max}$ - Maximum DTW, μ_{ρ} - mean velocity profile correlation

3.3 Simulating a Pedestrian-Vehicle Scenario

The openSCENARIO files, resulting from the reconstruction of the dedicated test drives were used as input for the simulation environment openPASS. An example for the simulated scenario is shown in Figure 13.

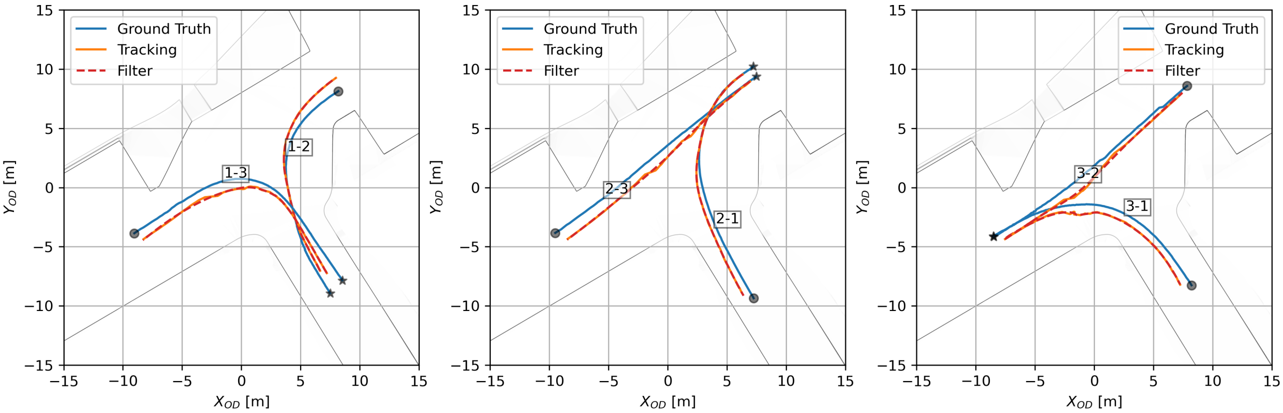

Figure 12. Qualitative comparison between the reconstructed trajectories (dashed line) and the geo-referenced GT motion data (solid lines). The left subfigure shows vehicle trajectories starting from road section 1, the middle from road section 2, and the right from road section 3, respectively.

4 Discussion

In this paper a workflow has been established to extract pedestrian-vehicle scenarios from camera-based observation system, suitable to the virtual assessment of ADAS.

4.1 Traffic Observation and Visual Perception Pipeline

Due to the internal service roads at the observation point chosen in this study there are frequent interactions between vehicles and pedestrians. However, at this particular point, accident with personal injuries have not been recorded in recent years, which can be explained mainly by the speed limit. The extension to observation points in public space as in [50, 4], would complement the scenario catalog as it results in other scenario configurations. Recordings of road sections with a higher speed limit, *i.e.* 50 km/h, regulated and unregulated crosswalks, or interactions with public transportation would be of additional value. In order to further quantify scenario relevance, it would be necessary to evaluate complexity and criticality based on common metrics such as traffic densities, or time to collision (TTC). Further observation points would be needed to underpin the results and to better understand intersection specific differences. The deliberate camera placement and viewpoint choices differs significantly from previous studies [49, 50, 4], and allows a higher level of detail which is useful to further increase the realism of the scenario description, especially with respect to road user interactions. In addition to the approach shown for to extracting road user trajectories automatically, the close-up observations will allow future work to investigate pedestrian attributes and even include realistic pedestrian postures [39] in simulation environments such as Car Learning to Act (CARLA) [10].

A realistic assessment of integrated safety systems should take into account the initial posture of a vulnerable road user (VRU) prior to a crash, as it can influence the accident kinematics and the resulting crash severities. Having the capability of reconstructing realistic postures of VRUs in critical situations, as described in [39, 26], and transferring them into a simulation environment could therefore enable more realistic virtual testing of integrated safety systems.

Furthermore, it should be noted that the presented trajectory reconstruction process could be applied to other observation points with relative ease. The adaptions required for this mainly concern two parts of the visual perception pipeline, namely tracking and ground plane projection and there in

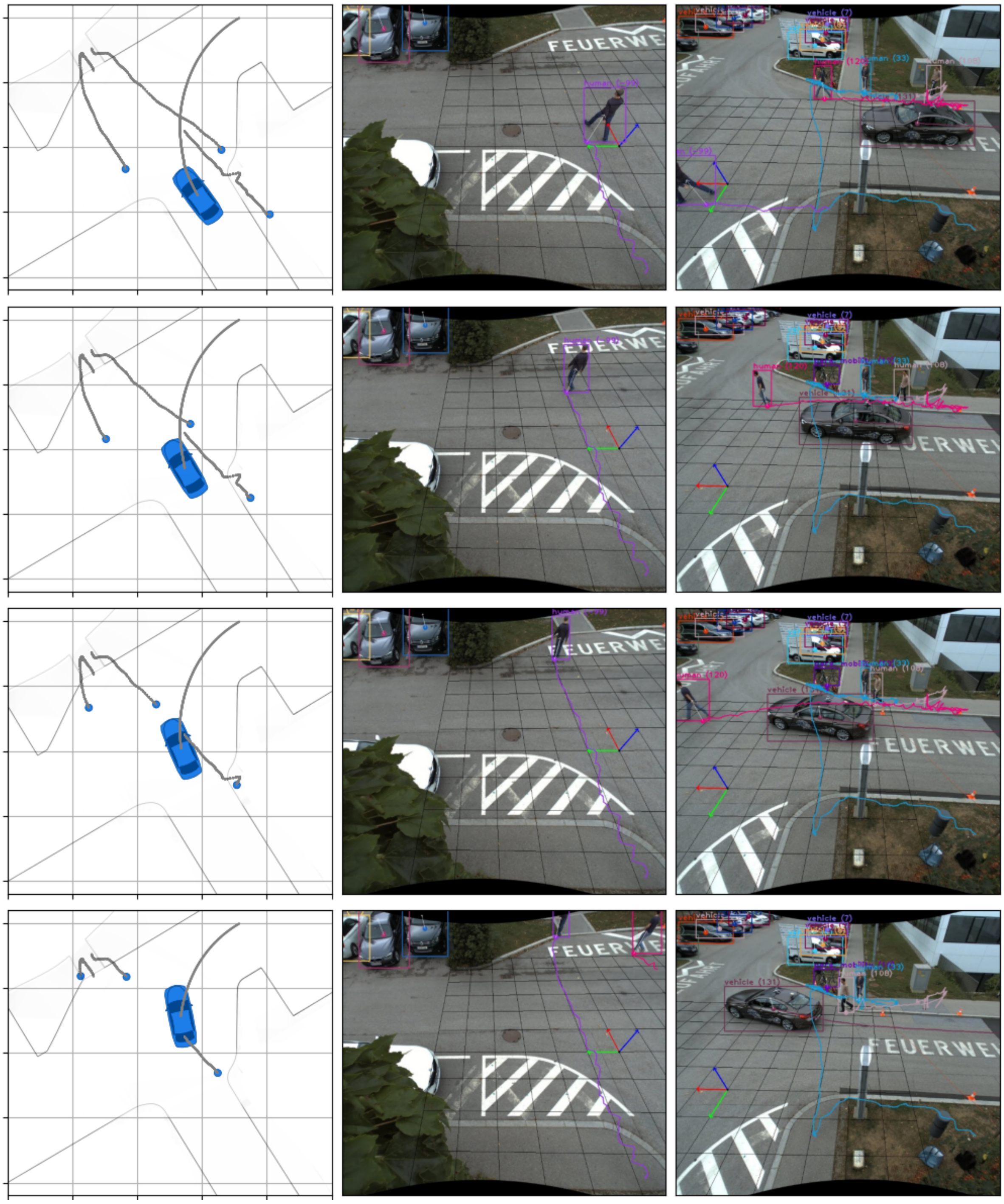

Figure 13. Visualisation of the scenario simulated in openPASS, alongside the corresponding video frames of the observation system, including the results of the MOT. Frames where taken at simulation time 0, 2, 4 and 6 seconds.

particular, the coordinate system alignment. The MOT algorithms can be adapted to other observation points, by adding additional training data and for the coordinate system alignment only one pair of corresponding GPS measurement and tag orientation/position measurement would be required to calculate the transformation T_{W2OD}.

4.2 Trajectory Reconstruction Accuracy

The performance of the trajectory reconstruction via perception pipeline, described in Section 2.4 depends on multiple factors. A central aspect of this is the correct operation of the MOT. The MOT performance shown in this paper is on state-of-the-art level, which could be quantified with common performance metrics. For the creation of a scenario catalog, the remaining misclassifications and ID switches play a minor role since they could be compensated in a post-processing step. A possible approach would be additional temporal and spatial sanity checks, *e.g.* checking if the start and end point of the trajectory are outside the observed area defined in Section 3.2.3. Overall, it can be assumed that the trajectories filtered in this way represent only a small subset of all reconstructed trajectories, and do not have a greater impact on the overall scenario distribution at this intersection. As expected, the highest reconstruction accuracy is reached near the center of each FOV. Furthermore, the road surface at the observation location in our study was highly uneven and thus, often violates the underlying assumptions of the central perspective projection model that the ground plane should be located at $z = 0$ for all locations in the observed image. The reconstruction accuracy is thus expected to improve notably when deployed at other observation points, where this assumption on the ground plane holds.

The investigation on vehicle trajectory accuracy estimation as presented in Section 3.2.3 shows a systematic deviation in the reconstructed path. The path reconstruction might be further improved by taking the distance to the image center into account, as well as the road network information, *i.e.* lane types, dedicated for specific road users. Apart from the general limits of reconstruction accuracy, as determined in Section 3.2.1, the representative vehicle point used for ground plane projection could possibly be enhanced, as shown by [44]. In our case, the representative point is a function of the bounding box size and is therefore of limited accuracy in the area where the vehicle is only partly visible, *i.e.* in the overlap of the cameras' FOVs. This transition affects both the reconstruction of the path as well as the reconstructed velocity. Since the path is currently reconstructed by spline interpolation, targeted smoothing by weighting the points with respect to the their WCS location could possibly enhance the results. For the velocity reconstruction, the effect is compensated by a mean filter, which could further be improved by using enhanced sensor fusion techniques.

Overall, the reconstructed trajectories were sufficiently accurate to permit realistic modelling as demonstrated by the low deviations in the specifically conducted accuracy estimation measurements and in the simulation.

4.3 Simulation

At the current stage only single scenarios were re-simulated. Nevertheless, the reconstructed trajectories could be lift the concrete scenario description to logical actions as shown in [35]. Logical scenarios, which are capable to model the entire traffic, can then be used for scenario-based assessment of ADAS, via stochastic simulations [40]. In general, traffic simulations include the dynamics and behavior of traffic participants (vehicles, pedestrians, *etc.*), the road network, environmental conditions (lighting, weather) and sensors like cameras, LiDARs, and Radio Detection and Ranging (Radars) [31]. The vehicle's sensors and the related perception algorithms for object detection and tracking provide essential input data for the ADAS that is being tested. Hence, modeling sensor capabilities and deficiencies with sufficient accuracy is thus a matter of the utmost importance for obtaining realistic simulation results. Different fidelity levels are required depending on which ADAS development process phase the sensor model is applied to. In early stages, where the focus lies on control or planning algorithm design, it is common to use models which provide object list outputs (see *e.g.* [27]), *i.e.* sensor and perception are encapsulated in one model. Later on, when also the in-vehicle perception software is tested, sensor models which provide raw data output (*e.g.* LiDAR point clouds [17]) are applied. An overview of various sensor model types and their underlying principles is given in [41].

5 Conclusions

The exemplary application of a newly developed workflow to bridge the gap between observed real-world pedestrian scenarios and scenarios in traffic simulations used for the assessment of active pedestrian protection systems was showcased within this paper. It was possible to simulate the observed scenarios with the simulation framework openPASS. The developed method and recorded data sets show great potential for future work and will support the development of more realistic virtual pedestrian scenarios and therefore a more realistic effectiveness assessment of ADAS in the future.

Acknowledgments This work was partially funded by the Austrian Research Promotion Agency (FFG) project INTERACT (879648). N. Kirillova also gratefully acknowledges the financial support provided by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association.

References

[1] Association for Standardization of Automation and Measuring Systems (ASAM). Openscenario®, 2022.

[2] Benfold, B., and Reid, I. Stable multi-target tracking in real-time surveillance video. In *CVPR 2011* (2011), pp. 3457–3464.

[3] Bernardin, K., and Stiefelhagen, R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *European Association for Signal Processing (EURASIP) Journal on Image and Video Processing* (2008).

[4] Bock, J., Krajewski, R., Moers, T., Runde, S., Vater, L., and Eckstein, L. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (2020), pp. 1929–1934.

[5] Bohannon, R. W. Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants. *Age and ageing 26*, 1 (1997), 15–19.

[6] Bouguet, J.-Y. Camera calibration toolbox for MATLAB., 2013.

[7] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 11618–11628.

[8] Detwiller, M., and Gabler, H. C. Potential Reduction in Pedestrian Collisions with an Autonomous Vehicle. In *The 25th ESV Conference Proceedings* (2017), NHTSA, Ed., ESV Conference Proceedings, NHTSA, pp. 1–8.

[9] Dobberstein, J., Bakker, J., Wang, L., Vogt, T., Düring, M., Stark, L., Gainey, J., Prahl, A., Mueller, R., and Blondelle, G. The Eclipse Working Group openPASS – an Open Source Approach to Safety Impact Assessment Via Simulation. In *The 25th ESV Conference Proceedings* (2017), NHTSA.

[10] Dosovitskiy, A., Ros, G., Codevilla, F., López, A., and Koltun, V. Carla: An open urban driving simulator, 2017.

[11] European Commission. *Road safety thematic report - Fatigue*. European Road Safety Observatory, Brussels, European Commission, Directorate General for Transport, 2021.

[12] Ferstl, D., Reinbacher, C., Riegler, G., Rüther, M., and Bischof, H. Learning Depth Calibration of Time-of-Flight Cameras. In *BMVC* (2015).

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

[14] GIS Steiermark. Airborne laserscanning-basierende höhendaten, 2021.

[15] GIS Steiermark. Digitaler atlas, 2022.

[16] Gruber, M., Kolk, H., Klug, C., Tomasch, E., Feist, F., Schneider, A., and Roth, F. The effect of P-AEB system parameters on the effectiveness for real world pedestrian accidents. In *The 26th ESV Conference Proceedings* (2019), NHTSA.

[17] Hanke, T., Schaermann, A., Geiger, M., Weiler, K., Hirsenkorn, N., Rauch, A., Schneider, S.-A., and Biebl, E. Generation and validation of virtual point cloud data for automated driving systems. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)* (2017), pp. 1–6.

[18] Hartley, R., and Zisserman, A. *Multiple View Geometry in Computer Vision*, 2 ed. Cambridge University Press, 2004.

[19] Jiang, K., Ling, F., Feng, Z., Ma, C., Kumfer, W., Shao, C., and Wang, K. Effects of mobile phone distraction on pedestrians' crossing behavior and visual attention allocation at a signalized intersection: An outdoor experimental study. *Accident Analysis & Prevention 115* (jun 2018), 170–177.

[20] Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, TaoXie, Kwon, Y., Michael, K., Changyu, L., Fang, J., V, A., Laughing, tkianai, yxNONG, Skalski, P., Hogan, A., Nadar, J., imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M. T., Marc, albinxavi, fatih, oleg, and wanghaoyang0106. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, oct 2021.

[21] Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering 82*, 1 (03 1960), 35–45.

[22] KALRA, N., AND PADDOCK, S. M. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice 94* (2016), 182–193.

[23] KOVACEVA, J., BÁLINT, A., SCHINDLER, R., AND SCHNEIDER, A. Safety benefit assessment of autonomous emergency braking and steering systems for the protection of cyclists and pedestrians based on a combination of computer simulation and real-world test results. *Accident Analysis & Prevention 136* (2020), 105352.

[24] LABBE, R. Kalman and bayesian filters in python, 2022.

[25] LI, Y., HUANG, C., AND NEVATIA, R. Learning to Associate: HybridBoosted Multi-Target Tracker for Crowded Scene. In *CVPR* (2009).

[26] LICH, T., MÖNNICH, J., SCHMIDT, D., AND VOSS, M. Preparation of an ai based real-time injury risk index estimation by deriving vru behavior from video-documented crashes. In *airbag 2022, 15th International Symposium and Exhibition on Sophisticated Car Safety Systems* (2022), Fraunhofer Institute for Chemical Technology ICT, pp. V20–19.

[27] LINNHOFF, C., ROSENBERGER, P., AND WINNER, H. Refining object-based lidar sensor modeling — challenging ray tracing as the magic bullet. *IEEE Sensors Journal 21*, 21 (2021), 24238–24245.

[28] LUITEN, J., OSEP, A., DENDORFER, P., TORR, P., GEIGER, A., LEAL-TAIXÉ, L., AND LEIBE, B. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision 129*, 2 (2020), 548–578.

[29] MATLAB. Roadrunner (r2022b), 2022.

[30] OXFORD TECHNICAL SOLUTIONS. User manual - rt3000 v3 and rt500 models, 2022.

[31] PAGE, Y., FAHRENKROG, F., FIORENTINO, A., GWEHENBERGER, J., HELMER, T., LINDMAN, M., OP DEN CAMP, O., VAN ROOIJ, L., PUCH, S., FRÄNZLE, M., SANDER, U., AND WIMMER, P. A Comprehensive and Harmonized Method for Assessing the Effectiveness of Advanced Driver Assistance Systems by Virtual Simulation: The P.E.A.R.S. Initiative. In *The 24th ESV Conference Proceedings* (2015), NHTSA.

[32] QGIS DEVELOPMENT TEAM. Qgis geographic information system, 2022.

[33] RAUCH, H. E., TUNG, F., AND STRIEBEL, C. T. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal 3*, 8 (1965), 1445–1450.

[34] REDMON, J., AND FARHADI, A. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 6517–6525.

[35] REICHENBÄCHER, C., RASCH, M., KAYATAS, Z., WIRTHMÜLLER, F., HIPP, J., DANG, T., AND BRINGMANN, O. Identifying scenarios in field data to enable validation of highly automated driving systems. In *Proceedings of the 8th International Conference on Vehicle Technology and Intelligent Transport Systems* (2022), SCITEPRESS - Science and Technology Publications.

[36] RICHARDSON, A., STROM, J., AND OLSON, E. AprilCal: Assisted and repeatable camera calibration. In *IROS* (2013).

[37] RISTANI, E., SOLERA, F., ZOU, R., CUCCHIARA, R., AND TOMASI, C. Performance Measures and a Dataset for Multi-Target, Multi-Camera Tracking. In *ECCV Workshop* (2016).

[38] ROSÉN, E., AND SANDER, U. Pedestrian fatality risk as a function of car impact speed. *Accident Analysis & Prevention 41*, 3 (2009), 536–542.

[39] SCHACHNER, M., SCHNEIDER, B., KLUG, C., AND SINZ, W. Extracting Quantitative Descriptions of Pedestrian Pre-crash Postures from Real-world AccidentVideos. In *2020 IRCOBI Conference Proceedings - International Research Council on the Biomechanics of Injury* (2020), IRCOBI, pp. 231–249.

[40] SCHACHNER, M., SINZ, W., THOMSON, R., AND KLUG, C. Development and evaluation of potential accident scenarios involving pedestrians and AEB-equipped vehicles to demonstrate the efficiency of an enhanced open-source simulation framework. *Accident Analysis & Prevention 148* (2020).

[41] SCHLAGER, B., MUCKENHUBER, S., SCHMIDT, S., HOLZER, H., ROTT, R., MAIER, F. M., SAAD, K., KIRCHENGAST, M., STETTINGER, G., WATZENIG, D., AND RUEBSAM, J. State-of-the-art sensor models for virtual testing of advanced driver assistance systems/autonomous driving functions. *SAE International Journal of Connected and Automated Vehicles 3*, 3 (oct 2020), 233–261.

[42] SCHWEBEL, D. C., STAVRINOS, D., BYINGTON, K. W., DAVIS, T., O'NEAL, E. E., AND DE JONG, D. Distraction and pedestrian safety: How talking on the phone, texting, and listening to music impact crossing the street. *Accident Analysis & Prevention 45* (mar 2012), 266–271.

[43] SEKACHEV, B., MANOVICH, N., ZHILTSOV, M., ZHAVORONKOV, A., KALININ, D., HOFF, B., TOSMANOV, KRUCHININ, D., ZANKEVICH, A., DMITRIYSIDNEV, MARKELOV, M., JOHANNES222, CHENUET, M., A ANDRE, TELENACHOS, MELNIKOV, A., KIM, J., ILOUZ, L., GLAZOV, N., PRIYA4607, TEHRANI, R., JEONG, S., SKUBRIEV, V., YONEKURA, S., VUGIA TRUONG, ZLIANG7, LIZHMING, AND TRUONG, T. opencv/cvat: v1.1.0, 2020.

[44] SEONG, S., SONG, J., YOON, D., KIM, J., AND CHOI, J. Determination of vehicle trajectory through optimization of vehicle bounding boxes using a convolutional neural network. *Sensors 19*, 19 (2019), 4263.

[45] SUN, P., KRETZSCHMAR, H., DOTIWALLA, X., CHOUARD, A., PATNAIK, V., TSUI, P., GUO, J., ZHOU, Y., CHAI, Y., CAINE, B., ET AL. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 2446–2454.

[46] TAO, Y., BOTH, A., SILVEIRA, R. I., BUCHIN, K., SIJBEN, S., PURVES, R. S., LAUBE, P., PENG, D., TOOHEY, K., AND DUCKHAM, M. A comparative analysis of trajectory similarity measures. *GIScience & Remote Sensing 58*, 5 (2021), 643–669.

[47] WANG, L., VOGT, T., DOBBERSTEIN, J., BAKKER, J., JUNG, O., HELMER, T., AND KATES, R. Multi-functional open-source simulation platform for development and functional validation of adas and automated driving. In *Fahrerassistenzsysteme 2016*. Springer, 2018, pp. 135–148.

[48] WOJKE, N., BEWLEY, A., AND PAULUS, D. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)* (2017), pp. 3645–3649.

[49] YANG, D., LI, L., REDMILL, K., AND ÖZGÜNER, U. Top-view Trajectories: A Pedestrian Dataset of Vehicle-Crowd Interaction from Controlled Experiments and Crowded Campus. In *2019 IEEE Intelligent Vehicles Symposium (IV)* (2019), pp. 899–904.

[50] ZHAN, W., SUN, L., WANG, D., SHI, H., CLAUSSE, A., NAUMANN, M., KÜMMERLE, J., KÖNIGSHOF, H., STILLER, C., DE LA FORTELLE, A., AND TOMIZUKA, M. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv:1910.03088 [cs, eess]* (2019).

[51] ZHOU, K., YANG, Y., CAVALLARO, A., AND XIANG, T. Omni-scale feature learning for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 3701–3711.