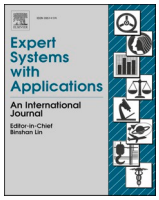
Expert Systems With Applications 304 (2026) 130821

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Knowledge PET3D: An interpretable framework for 3D near-miss detection in thermal traffic video

Arnd Pettirsch, Alvaro Garcia-Hernandez *

Institute of Highway Engineering, RWTH Aachen University, 52074 Aachen, Germany

ARTICLE INFO

Keywords:
Knowledge-driven traffic safety
Near-miss detection
3D trajectory analysis
Rule-based conflict interpretation
Anomaly detection
Engineering decision support

ABSTRACT

Traffic safety engineers often rely on retrospective crash data, limiting their ability to proactively identify systemic risks in road environments. To address this gap, this work presents Knowledge PET3D, a novel, privacy-preserving framework that enables automatic long-term traffic observation, detects safety-relevant interactions, and delivers interpretable video snippets to support informed engineering decisions. The system integrates monocular 3D detection, maneuver-specific rule modeling, and transformer-based anomaly detection to identify near-miss events from thermal video. The system filters non-informative interactions based on rule compliance and behavioral response, enabling interpretable conflict sets suitable for manual review. Compared to conventional PET2D and PET3D baselines, Knowledge PET3D achieves over 4x more true positives and reduces false positives by up to 93%. It delivers high precision across varied urban contexts (22% at a complex signalized intersection and up to 75% at a simpler yield-controlled site), while keeping conflict volumes verifiable by humans. The framework further achieves 87.0% correct classified maneuvers, 92.9% clustering accuracy, and 96.5% correct interpreted traffic rules. Knowledge PET3D advances traffic safety diagnostics by uncovering latent risks while maintaining engineering interpretability and operational scalability.

1. Introduction

Road safety audits (RSAs) are a core practice in transport infrastructure planning across many regions, including the European Union (Directive, 2008), the United States (Administration, 2006), and Asia (CAREC Road Safety Engineering Manual 1, 2018). They provide a structured, expert-led evaluation of road environments using a combination of historical crash data, traffic volumes, road geometry, traffic control features and other criteria (Othman and Ros, 2013). However, such audits often rely on retrospective data and high-level heuristics, limiting their ability to proactively identify systemic safety risks before crashes occur. This leads to a broader challenge in safety engineering: How to automatically observe traffic in longer periods to detect safety relevant situations and present them in an interpretable way to support traffic safety engineers' decision making.

The traffic safety community has increasingly turned to surrogate measures such as traffic conflicts and near-misses to enable proactive risk assessment. Those events occur more frequently and capture latent risk before actual crashes happen (Hydén, 1987). Many studies use surrogate safety measures like Time to Collision (TTC) or Post Encroachment Time (PET) which estimate risk using spatial and temporal proximity between road users (Singh and Das, 2021). While those methods remain prominent in recent studies like Mukherjee et al. (2025) (Mukherjee, 2025), they produce large volumes of data (for example about 717 s PETs below 1 s from 4871 vehicles (Islam et al., June 2023) that are difficult for engineers to interpret and act upon.

To overcome these issues, historically Tarko et al. define conflicts by evasive maneuvers (Tarko et al., 2009) for example identified by deviations from reference trajectories of unhindered road users (Johnsson and Laureshyn, 2022). Yet, such methods often rely on fixed spatial thresholds or specific road-user classes (Kar et al., 2024), limiting their generality and scalability in real-world traffic environments (Johnsson and Laureshyn, 2022). Recent full AI-based methods, such as the deep transfer learning framework by Hou et al. (2025) (Hou et al., May 2025), predict conflict likelihood directly from vehicle trajectories using transformer networks. While these models improve performance and adaptability, they require large labeled datasets and lack interpretability, which makes them difficult to trust in engineering workflows where transparency and context are critical (Mannering et al., March 2020).

* Corresponding author.
E-mail address: alvaro@isac.rwth-aachen.de (A. Garcia-Hernandez).

https://doi.org/10.1016/j.eswa.2025.130821
Received 28 July 2025; Received in revised form 9 November 2025; Accepted 12 December 2025
Available online 13 December 2025

Effective conflict detection systems in engineering practice must go beyond abstract proximity metrics or opaque AI predictions. As highlighted by Mannering et al. (Mannering et al., March 2020) interpretability is essential in traffic safety modeling, where understanding whether a predicted conflict results from random behavior or a systemic design flaw is critical for informed engineering decisions. This implies that conflict detection should not be viewed as the final decision-making step, instead, its primary role is to filter and prioritize events for human review. Since engineering interventions cannot rely solely on numerical indicators, systems must produce outputs that are explainable and contextually meaningful to practitioners.

Therefore, the choice of traffic sensor is crucial. Among sensing options like LiDAR and radar (Tasgaonkar et al., 2020) video cameras offer a unique advantage: they enable human-interpretable visual context. Short video snippets allow engineers to understand whether sightlines were obstructed, whether a road user was not seen, or whether the incident was due to random behavior or a systemic design flaw. Despite the advantages of interpretable video snippets, conventional RGB cameras raise significant concerns around privacy and robustness. Most countries restrict the recording of identifiable personal data (Supervisor, 2023), and visible-light cameras often fail in low-light or adverse weather conditions (Bhadoriya et al., 2022). Thermal imaging addresses both limitations. It captures infrared radiation, naturally anonymizing scenes by hiding facial and license plate details, and is robust to lighting and weather conditions (Alldieck et al., 1947). This makes thermal video ideal for privacy-compliant, 24/7 traffic monitoring. Nevertheless, most existing camera-based traffic conflict detection systems face key limitations. Many rely on simplified point-based projections into world coordinates, ignoring object dimensions and orientations (Wang et al., 2025). Others focus only on specific object classes and geometric simplifications (Abdel-Aty et al., 2022), or depend on manual or semi-automatic data collection, which limits scalability due to the high workload (Wei et al., February 2019). These constraints reduce both the accuracy and applicability of current camera-based conflict analysis in complex, real-world settings.

Although progress continues, existing approaches to traffic conflict detection remain fragmented and limited in scope. They often focus on specific object classes, rely solely on handcrafted thresholds, or depend on manual data collection, which limits scalability and generalization. Crucially, no existing work systematically relates detected conflicts to applicable traffic rules and expected road user behavior in context while also delivering interpretable video snippets to support engineering judgment. As a result, there is a clear need for an integrated system that combines privacy-compliant sensing, precise trajectory extraction, traffic rule interpretation, and conflict analysis in a unified and automated pipeline suitable for engineering applications.

This work presents the first complete framework for privacy-preserving near-miss detection using thermal roadside video, enabling scalable, interpretable, and privacy-compliant traffic safety monitoring. The system integrates thermal 3D tracking, transformer-based anomaly detection aligned with geometric clusters, and formalized rule evaluation to produce interpretable results. Within this framework, Knowledge PET3D is introduced, incorporating behavioral cues and traffic rule context to identify meaningful near-miss events. By filtering interactions through this knowledge-based lens, the framework delivers focused and human-interpretable conflict reports, including video snippets, that support proactive safety audits and infrastructure evaluation. Evaluated on 272 h of thermal video from two urban locations, the framework demonstrates strong performance and practical value by translating raw sensor data into structured, rule-informed outputs that advance knowledge-driven traffic safety analysis.

2. Methodology

2.1. Problem description

Proactive near-miss detection in traffic involves the integration of sensing, modeling, interpretation, and reporting. The main challenge is the transformation of raw visual input Θ into interpretable knowledge that aids traffic safety decision-making.

A structured transformation pipeline Λ is proposed (1), converting sensor input Θ into 3D object trajectories T, identifying spatiotemporal interactions I, and generating semantically rich reports P that include visual and contextual information such as traffic rules. This layered process is defined as:

$$\Lambda : \Theta \rightarrow \mathrm{T} \rightarrow \mathrm{I} \rightarrow \mathrm{P} \tag{1}$$

2.2. Proposed framework

This work introduces Knowledge PET3D, a refined conflict definition that builds on classical PET (Allen et al., 1978). It uses full 3D object volumes and a temporal window of 3 s, supported by Peesapati et al. (Peesapati et al., January 2013), who found that PETs below 3 s are commonly associated with potential crash scenarios. This threshold offers a practical basis for capturing a wide range of interactions, for later filtering. As illustrated in Fig. 1, a conflict is confirmed only when three conditions are met: a PET-based interaction, a traffic rule violation, and anomalous behavior by the road user whose right of way was infringed. Anomalies are flagged using a Local Outlier Factor (LOF) threshold of 1.5 to ensure the presence of atypical motion. For each such event, the system generates a structured summary including maneuver type, rule status, PET value, LOF anomaly score, and a thermal video snippet. These reports enable interpretable, engineering-friendly conflict analysis and support prioritization by severity.

Fig. 2 illustrates the architecture of the proposed framework, which is structured into three processing layers, each serving a distinct role. The video processing layer extracts object trajectories from thermal video through detection, tracking, and refinement, providing accurate, privacy-compliant motion data as a foundation. The conflict detection layer prepares this data for safety interpretation by identifying spatiotemporal interactions, clustering common movement paths, and detecting anomalies indicative of evasive behavior. The engineering interpretation layer applies rule and maneuver classification to filter relevant cases and generate structured reports, supporting expert reasoning and decision-making.

2.3. video processing layer

2.3.1. Thermal imaging and privacy

Thermal video is recorded using AXIS Q1952-E (Communications, 2021) cameras at 640 × 480 resolution and 30 fps. Pixel intensities encode relative heat emissions, enabling scene interpretation while ensuring anonymization for privacy-preserving traffic monitoring. While thermal video is used here for detection due to its privacy-preserving properties, all subsequent components operate on georeferenced 3D detections and are modality-agnostic. The pipeline can equally process inputs from RGB, LiDAR, or multimodal fusion systems as long as 3D positions and object attributes are available. Thermal imagery remains advantageous for privacy-compliant storage of interpretable snippets that support engineering review.

2.3.2. Georeferenced transformation

Detections are projected from image to world coordinates to enable metric reasoning about position, dimension, and speed. This includes lens distortion correction, application of calibrated camera intrinsics, and ray-plane projection. The whole process follows (Pettirsch and Garcia-Hernandez, April 2025).

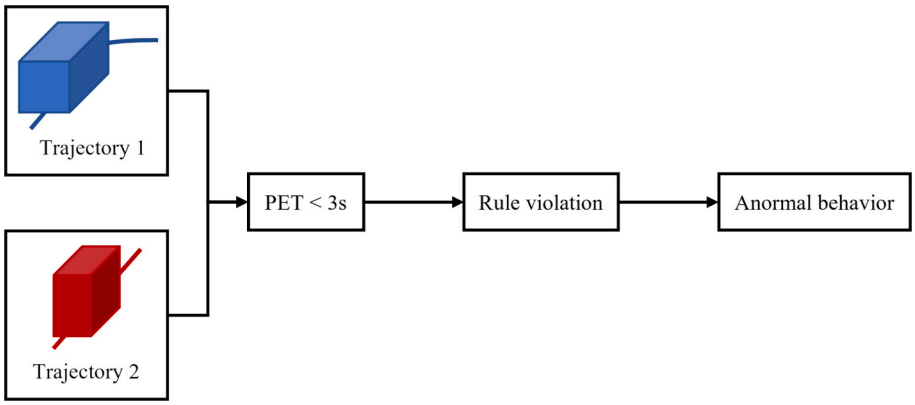


Fig. 1. Knowledge PET3D: A conflict is detected when two road users interact with short temporal spacing, a rule violation is present, and the disadvantaged party shows anomalous behavior.

2.3.3. Detection and 3D localization

A retrained version of the ProjNet-Model (Pettirsch and Garcia-Hernandez, April 2025) performs monocular 3D detection, predicting 2D bounding boxes, classes, keypoints, 3D box dimensions, and orientations. The model supports seven classes: motorcycle, car, truck, bus, pedestrian, bicycle, and e-scooter. To reduce false positives, manually defined detection zones (~ 40 m radius) are used per location. Only detections above a 0.25 confidence threshold within these zones are retained.

2.3.4. Tracking and matching

2.3.4.1. Trajectory estimation in world coordinates. The method operates directly in georeferenced 3D coordinates, using a Kalman filter (Kalman, 1960) with a constant-velocity motion model (Weng et al., 2020). Each object is represented by a state vector S (2) comprising 3D position (x,y,z), dimensions (l,w,h) and velocity components (v_x,v_y,v_z).

$$S = (x, y, z, l, w, h, v_x, v_y, v_z) \tag{2}$$

Angular velocity and acceleration are excluded for noise robustness. Observations update the state via position and dimension estimates.

2.3.4.2. Detection-to-Track Association. Detection-to-track association is based on a cost function similar to the one used in EagerMOT (Kim et al., 2021), which combines geometric distance, box similarity, and orientation penalty. In addition, a class-aware consistency term $P_{i,j}$, balanced by the factor w, it introduced to improve robustness in dense traffic. For detection and track positions p_i,p_j, dimensions d_i,d_j, and yaws Θ_i, Θ_j, the cost it:

$$C_{i,j} = \left(2\|p_i - p_j\|_2 + \|d_i - d_j\|_2\right) \bullet \left(2 - \cos(\Theta_i - \Theta_j)\right) \bullet \left(1 + wP_{i,j}\right) \tag{3}$$

$w = 0.5$ is a weighting factor, $P_{i,j}$ encodes semantic consistency, defined in (4) with s_i the detection confidence and the Kronecker delta δ with C_i, C_j the detection and track class. This function penalizes confident class mismatches while tolerating low-confidence uncertainty. The value $w = 0.5$ was chosen to give geometric and semantic terms approximately equal influence, given their similar scale. This balance was found to be stable across various traffic densities without fine-tuning. For each frame, a cost matrix between detections and existing tracks is computed using (3). Greedy bipartite matching is then applied to this matrix, with associations accepted only if the corresponding cost is below a class-specific threshold ($t_{Motorcylce} = 7.5, t_{Car} = t_{Truck} = t_{Bus} = 9, t_{Person} = 2.5, t_{Bicylce} = t_{E-Scooter} = 5$).

$$P_{i,j} = \left(1 - \delta_{C_i,C_j}\right)s_i + \delta_{C_i,C_j}(1 - s_i) \tag{4}$$

2.3.4.3. Robustness and lifecycle Strategy. To suppress artifacts during close encounters, yaw correction if applied. If the yaw deviation exceeds $\pi/2$, a π-rotation aligns orientation. This avoids unrealistic 180° flips and stabilizes trajectories. Tracks are initiated after three consistent unmatched detections and terminated after 10 missed frames or exiting the detection zone for 10 frames. This suppresses false positives and outdated trajectories.

2.3.5. Trajectory enhancement

A multi-stage pipeline (Fig. 3) comprising initial filtering, smoothing, splitting, matching, and final filtering with attribute voting is applied.

2.3.5.1. Initial Filtering. Tracks with < 50 % frame coverage or duration < 1 s are discarded to remove unstable data.

2.3.5.2. Smoothing and Kinematic Estimation. Smoothing is applied before splitting and after matching. Object positions $p = (x, y, z)$, are smoothed independently along each axis using a zero-padded moving average with kernel size $k = 5$ (5). Velocities and accelerations are computed via central differences as shown in (6) and (7) with $\Delta t = 1/fps$.

$$\hat{p}_t = \frac{1}{2k+1}\sum_{i=-k}^{k} p_{t+1} \tag{5}$$

$$v_t = \frac{\hat{p}_{t+1} - \hat{p}_{t-1}}{2\Delta t} \tag{6}$$

$$a_t = \frac{\hat{p}_{t+1} - 2\hat{p}_t + 2\hat{p}_{t-1}}{(\Delta t)^2} \tag{7}$$

Following (Rezaei et al., October 2023) yaw is derived from smoothed velocity vectors using (8). Circular smoothing using unit complex averaging ensures continuity (9).

$$\Theta = atan2(v_y, v_x) \tag{8}$$

$$\hat{\Theta} = \arg\left(\frac{1}{2k+1}\sum_{i=-k}^{i=k} e^{i\Theta_{t+1}}\right) \tag{9}$$

2.3.5.3. Trajectory splitting and matching. Tracks are segmented at discontinuities, defined as yaw changes > $\pi/4$ or spatial jumps > $2m$. This captures occlusions, reidentifications, and errors, ensuring subsegments reflect coherent motion. Post-hoc linking reconnects fragmented segments likely from the same object. Matching uses a cost matrix based on endpoint − start point proximity, class, and orientation using (3).

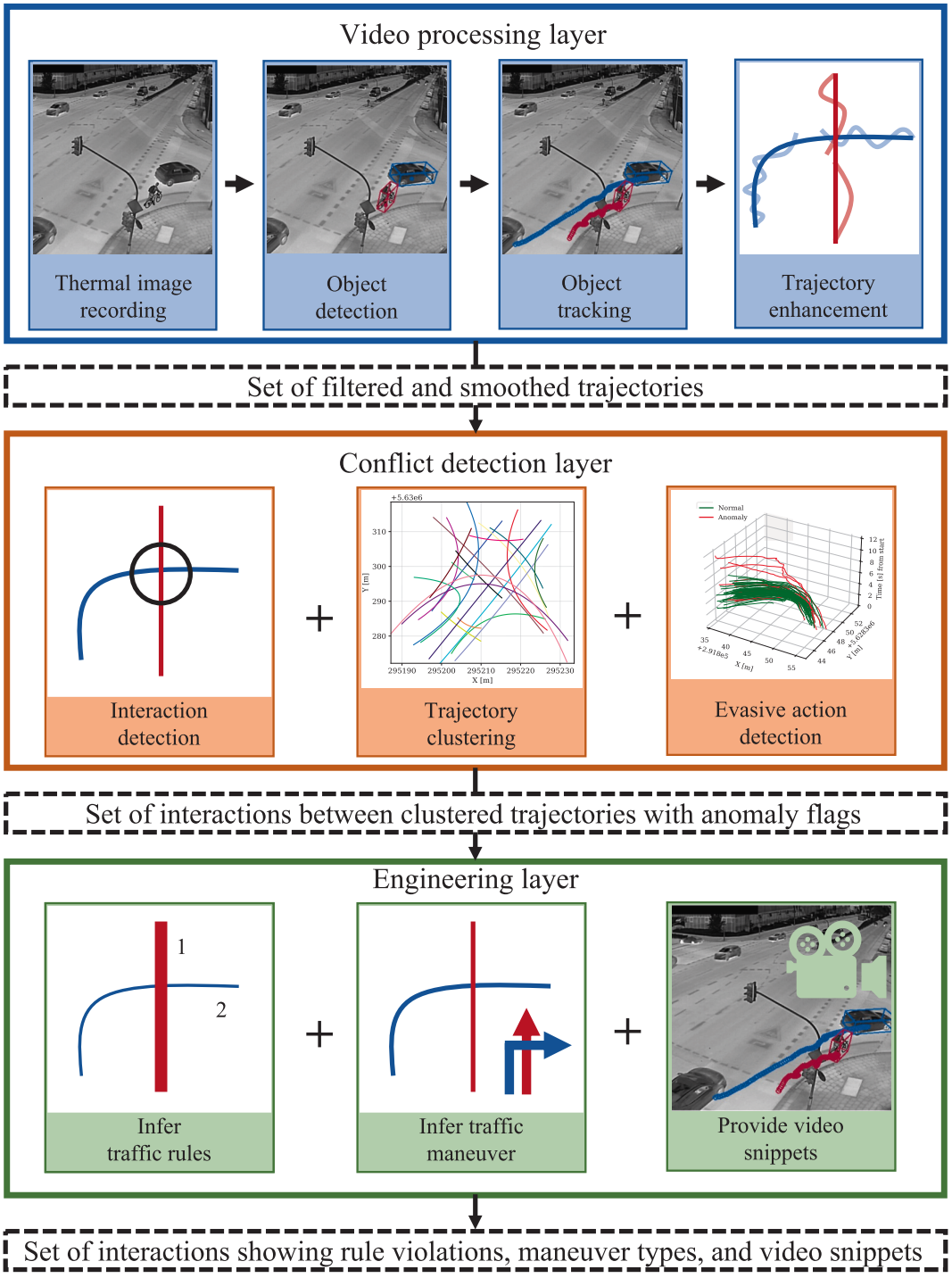


Fig. 2. Overview of the proposed framework showing the different processing layer. From raw sensor data processing in video processing layer, to conflict detection layer and semantic meaningful engineering layer as decision support.

Greedy bipartite matching yields continuation pairs and unmatched residuals.

2.3.5.4. Fine Filtering and attribute voting. Trajectories must be active ≥ 50 % of frames, meet duration (2 s) and distance thresholds (5 m), and show no extreme yaw fluctuations (>3.2 rad). To remove unstable behavior at the beginning and end of tracks, the first and last 10 data points are discarded. Final class labels are assigned by majority voting (Pettirsch and Garcia-Hernandez, April 2025), while object dimensions are computed as temporal averages over valid detections.

2.4. Conflict detection layer

2.4.1. Detect interactions

PET (Allen et al., 1978) is used to prefilter interactions. It measures the time gap between one object exiting and another entering a shared 3D space:

$$PET = t_{entry,2} - t_{exit,1} \tag{10}$$

IoU is computed between the bottom areas of 3D boxes, an IoU > 0.1 is considered as an intersection, accounting for minor inaccuracies when

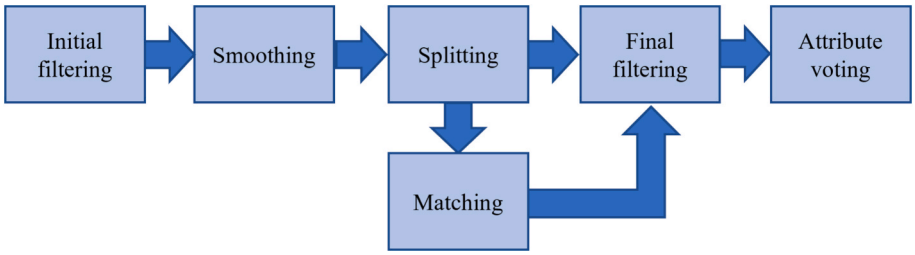


Fig. 3. Overview trajectory enhancement. After initial filtering and smoothing, unrealistic trajectories are split and matched if possible. Finally, there is filtering and attribute voting.

vehicles are close. For each vehicle pair, the minimum PET with overlap is recorded (Laureshyn et al., 2010).

2.4.2. Trajectory clustering

To capture the most common paths through the intersection, spatial clustering is performed purely on geometric shape in BEV space, abstracting away detailed trajectory and temporal variations. This provides a general path structure for later anomaly detection to focus on behavioral deviations.

2.4.2.1. Cluster definition. Trajectory data from the first 24 h of video recording was used to define clusters, as dominant movement patterns typically stabilize within a full weekday of observation, covering regular morning and evening peaks. Each trajectory is downsampled to 5 equally spaced keypoints (0 %, 25 %, 50 %, 75 %, 100 % of traveled distance) to capture global path shape while suppressing noise. Pairwise Euclidean distances across all trajectories and keypoints form a 3D distance tensor, which is collapsed to a 2D matrix using the maximum distance per trajectory pair. DBSCAN (Ester et al., 1996) is applied to this matrix with a maximum reachable distance of 0.5 and a minimum of 5 trajectories per cluster to identify path-shape clusters.

Each cluster's mean trajectory is computed by averaging member keypoints, then smoothed by fitting second-degree polynomials independently to $x(t)$, $y(t)$ and $z(t)$ over normalized time $t \in [0, 4]$. Clusters with similar shapes are merged iteratively using combined distance criteria: a maximum point-to-curve distance < 1 m or a total keypoint distance < 5 m.

2.4.2.2. Cluster assignment. New trajectories are downsampled the same way and compared to cluster mean curves using point-to-curve distances. A trajectory is assigned to the best-matching cluster if either the sum of all keypoint distances is below 10 m or the maximum distance is below 3 m. Unmatched trajectories remain unassigned. To account for direction, the starting point is compared to both ends of the cluster curve. If it aligns more closely with the reverse direction, the trajectory is assigned a negative cluster ID to indicate reversed movement.

2.4.3. Evasive action detection

To detect evasive actions such as deceleration, acceleration, swerving, or abrupt braking, each trajectory is uniformly encoded into a fixed-length sequence of displacement vectors $(\Delta x, \Delta y)$, capturing short-term motion patterns over time. An encoder–decoder architecture (Vaswani et al., 2017), as shown in Fig. 4, processes these sequences. The encoder receives trajectories corrupted via a combination of masking and noise injection: approximately 40 % of the displacement vectors are set to NaN in randomly placed contiguous segments of up to 5 steps, while the remaining unmasked points are perturbed with Gaussian noise ($\sigma = 0.1$). The encoder maps this corrupted input into a latent space, while the decoder reconstructs the original sequence. An auxiliary classification head is appended to the encoder to encourage the latent representation to encode motion semantics by predicting the vehicle class.

The model consists of a Transformer encoder with 4 layers, each having 4 attention heads and a hidden dimension of 128. The input is a sequence of 3D vectors $(\Delta x, \Delta y, maskflag)$, which are embedded to 128 dimensions and enriched with learned positional encodings (for sequences up to 1200 steps). The output latent space preserves this 128-dimensional embedding per time step. The full model includes a linear decoder for trajectory reconstruction and an auxiliary classification head that predicts the object class from pooled encoder features. In total, the network contains approximately 980,000 trainable parameters. To ensure generalization across varying camera viewpoints, trajectories are first rotated according to their initial yaw and shifted to the origin before displacement vectors are computed. (Vaswani et al., 2017).

2.4.3.1. Self-supervised training. The model is trained using a self-supervised objective: it must reconstruct the original displacement sequence from the masked and noise-augmented input. Additionally, the model predicts the object class as an auxiliary task. The ground-truth class is denoted by c, and $\widehat{p_k}$ denoted the predicted probability of class k. The total loss combines mean squared error over the masked and unpadded displacement vectors (first term in (11)), and a cross-entropy classification loss (12) for the predicted class distribution, weighted by a factor $\lambda = 0.3$:

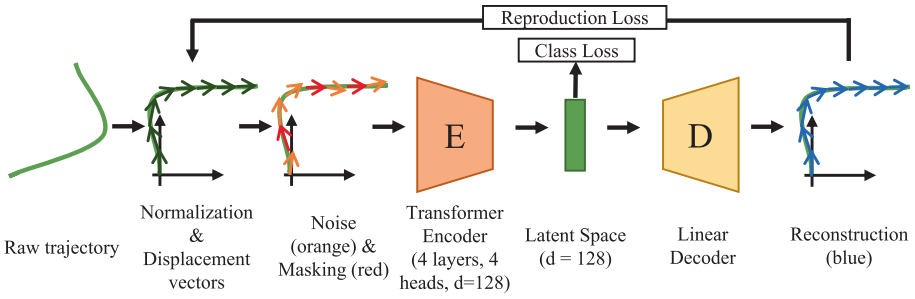


Fig. 4. Overview of the encoder decoder structure to learn vectorized representation of the trajectories.

$$L = \| \begin{pmatrix} x_i - \widehat{x_i} \\ y_i - \widehat{y_i} \end{pmatrix} \|_2 + \lambda L_{class}(c, \widehat{c}) \tag{11}$$

$$L_{class}(c, \widehat{c}) = -\sum_{k=1}^{K} 1_{[c=k]} log \widehat{p_k} \tag{12}$$

For comparative evaluation, an LSTM-based encoder with a similar embedding dimension and classification head was trained under identical conditions. This allows assessing whether the transformer's attention mechanism provides any measurable advantage in anomaly detection performance.

2.4.3.2. Latent space-based anomaly detection. Due to recording noise and human variability, many trajectories exhibit irregularities. To minimize false positives, anomaly detection is restricted to trajectories that are involved in an interaction and belong to the same cluster and object class. When possible (number of remaining trajectories ≥ 20), detection is further limited to conflicts with rule violations. Each trajectory is mapped to a latent vector $z = E(\tau) \in R^d$ using the trained encoder. For each cluster–cluster conflict pair, Local Outlier Factor (LOF) (15) (Breunig et al., 2000) is computed cluster-wise using cosine distance and $k = 20$ neighbors. LOF compares the Local Reachability Density (LRD) (14) around a trajectory to that of its neighbors. For each neighbor $z_i \in N_k(z)$, the reachability distance is defined as (Breunig et al., 2000):

$$rd_k(z, z_i) = max(dist(z, z_i), dist(z, z_k)) \tag{13}$$

$$LRD(z) = \left(\frac{1}{k} \sum_{z_i \in N_k(z)} rdk(z, z_i) \right)^{-1} \tag{14}$$

LOF values near 1 indicate typical behavior, while higher values reflect outliers such as evasive actions. A threshold of 1.5 is used to classify anomalies. This value was adopted without further hyperparameter tuning, as it provided stable and interpretable results across both study locations. While the original LOF formulation (Breunig et al., 2000) does not specify a fixed threshold, values above 1.5 are widely considered indicative of anomalous behavior.

$$LOF(z) = \frac{1}{k} \sum_{z_i \in N_k(z)} \frac{LRD(z_i)}{LRD(z)} \tag{15}$$

2.5. Engineering layer

An engineering knowledge layer integrates traffic rules and maneuver classifications to contextualize interactions, enabling the system to distinguish between compliant behavior and potential conflicts.

2.5.1. Traffic maneuver classification

The framework assigns conflict types $m \in M$ by analyzing geometric relationships between mean trajectories of clustered object movements, reducing sensitivity to noise.

If trajectories intersect and start points are within 2 m of the means, the conflict is labeled 'Diverging'. If the start distance exceeds 2 m, heading changes determine the type: deviations over 30° imply a 'Turn-across-path' (left/right), while stable headings indicate 'Crossing'.

Without intersection, start–end point proximity is checked. If only one is below 2 m, increasing or decreasing distance indicates 'Diverging' or 'Merging', respectively. Parallel cases use heading alignment: deviations $< 20°$ suggest 'Following', $> 160°$ imply 'Head-on'. If none apply, mean angular deviation over $k = 20$ frames (16), before the interaction is used for classification as shown in (17).

$$\Delta\Theta = \frac{1}{k} \sum_{i=0}^{i=k} \Theta_{1,k-i} - \Theta_{2,k-i} \tag{16}$$

$$m = \left\{ \begin{array}{c} crossing if \Delta\Theta > 45^{\circ} \\ merging if 15^{\circ} < \Delta\Theta \geq 45^{\circ} \\ following if \Delta\Theta \leq 15^{\circ} \end{array} \right\} \tag{17}$$

2.5.2. Traffic rule interpretation

Traffic rules are encoded in a structured knowledge base $K = \{r_1, r_2, \cdots, r_n\}$, where each rule r describes which cluster is expected to have the right-of-way in a given interaction type. The rule is assumed as violated if the object with lower priority enters the interaction zone first:

$$r(\tau_1, \tau_2) = violated if priority(C_2) > priority(C_1) but \tau_1 enters first. \tag{18}$$

In this work, only the right-of-way rule is inferred, as it provides a general and interpretable basis for identifying conflicts. More detailed rules (e.g., STOP signs or location-specific regulations) are context-dependent and are therefore left to engineering judgment in the subsequent safety assessment.

The rule check is only applied when the priority confidence of the higher-priority cluster exceeds 0.66, ensuring reliable interpretation. While no cluster pairs with fewer than 10 interactions occurred in our dataset, such cases can be excluded in future applications to maintain robustness in low-data scenarios. For Following and Diverging maneuvers, or when priority is uncertain, the rule flag is always set to true to avoid overreporting of non-relevant cases. In settings with known regulatory metadata (for example, signal plans or yield control), the legal rule can be supplied as a fixed prior and the behavior-derived estimate is reported only as an auxiliary check.

The system automatically derives traffic rules from real-world behavior using the clusters and interaction detection described above. Based on the assumption that drivers generally follow the traffic rules the priorities are assigned on the frequencies of arrivals at the conflict zone in the n interactions between trajectories of the clusters:

$$priority(C_i)_{C_j} = \frac{\sum_{i=0}^{i=n} 1 if \tau_i \in C_i enters first}{n} \tag{19}$$

3. Experimental design and data collection

3.1. Observation locations

To evaluate the framework in varied urban conditions, two distinct junctions were selected, each with unique geometric and regulatory characteristics. Both locations are situated within the same mid-sized European city. Approximately one week of thermal video was recorded per site (142.5 h and 129.3 h), providing representative data. Fig. 5 shows example thermal and satellite images of both locations.

3.1.1. Location A – signalized intersection

This four-arm signalized junction connects a major arterial road (Street A) with a smaller side road (Street B), handling substantial multimodal traffic. Of particular interest are simultaneous green phases, which create overlapping movement patterns and frequent conflict points. Fig. 6 summarizes the average number and composition of detected road users per day over one week of observation. On average, approximately 16,400 road users were recorded daily, predominantly motor vehicles (≈ 14,700 cars), along with ≈ 1100 bicycles, 400 pedestrians, and smaller shares of motorcycles, buses, trucks, and e-scooters. This distribution highlights the multimodal character of the site and provides a quantitative baseline for future comparative studies.

3.1.2. Location B – right-turn lane

This unsignalized right-turn slip lane connects Street C to Street D, where vehicles cross a shared pedestrian and bicycle path. Drivers are required to yield to crossing road users. A permanent flashing amber light provides general warning at the crossing. Fig. 7 shows the corresponding vehicle distribution at Location B. Daily traffic volumes

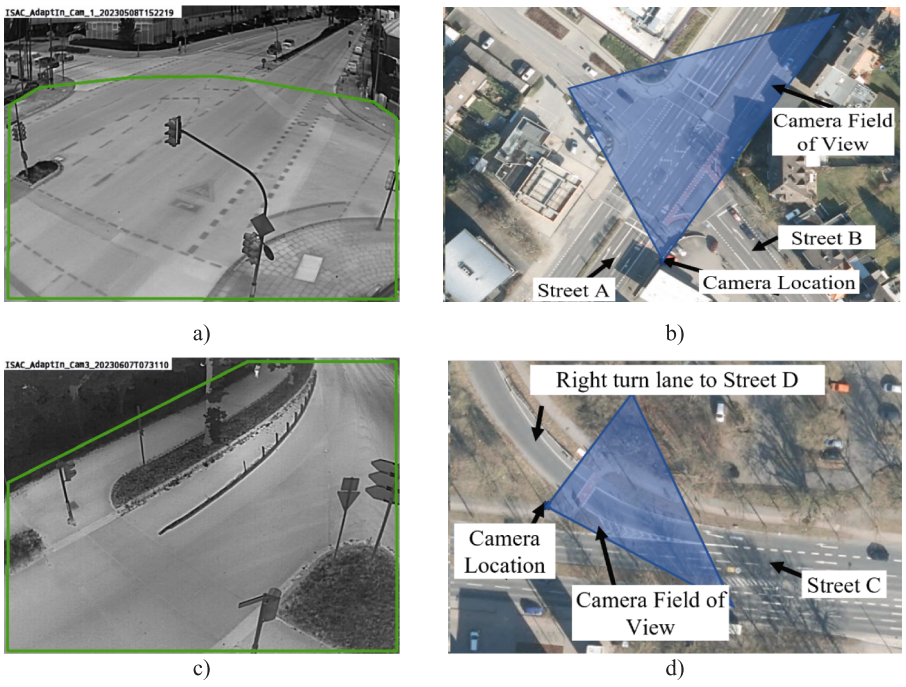


Fig. 5. A), b) intersection a– Signalized intersection. a) Example image from thermal camera with detection zone in green b) Blue shows the camera field of view c), d) Intersection B – Right turn lane c) Detection zone in green on thermal example image d) The camera field of view in blue. All Satellite images from (Esri, 2025).

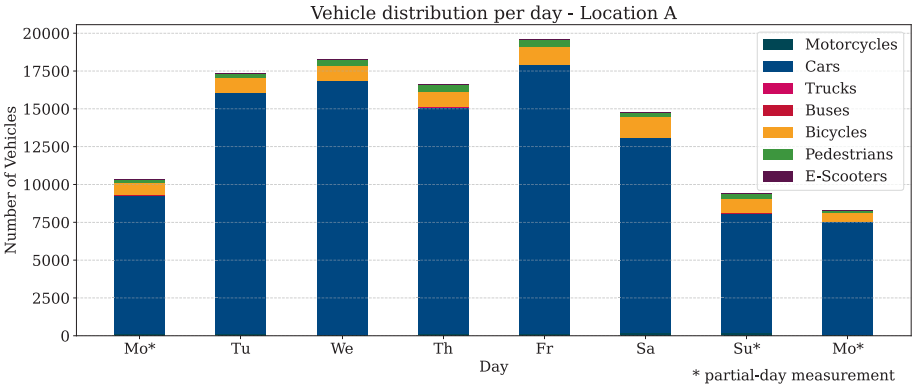


Fig. 6. Vehicle distribution per day at Location A.

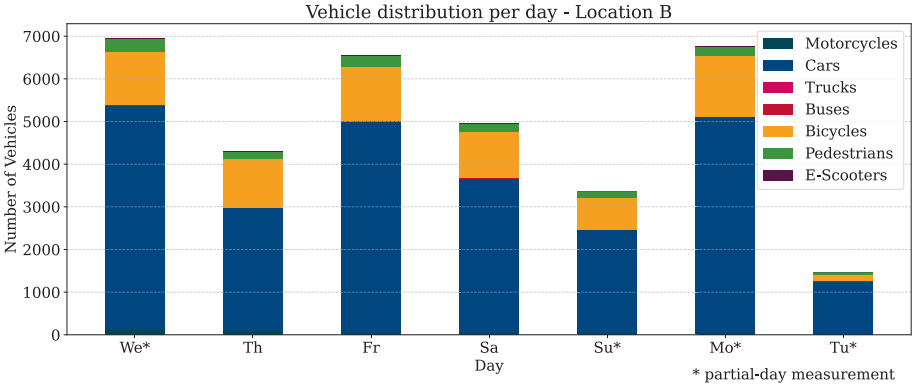


Fig. 7. Vehicle distribution per day at Location B.

average ≈ 4900 road users, composed of ≈ 3600 cars, 1000 bicycles, 190 pedestrians, and small numbers of motorcycles and e-scooters. Compared with Location A, the higher share of cyclists and pedestrians makes this site particularly relevant for vulnerable-road-user interactions.

3.2. Evaluation metrics

The framework was evaluated for its ability to reliably detect relevant near-miss events, minimize false positives, and support interpretability for traffic engineering.

3.2.1. Conflict detection evaluation

Precision (P) was computed as the ratio of true positives (TP) to all detected events ($TP + FP$), based on manual verification of all detections (20). False positives were defined as detections lacking any observable critical interaction, rule violation, or evasive behavior. All labels were created by a single expert based on visual inspection of thermal video and trajectory data. While this ensured consistency, it may introduce subjectivity. Double-coding with multiple raters is planned in future work. For statistical testing, we used Fisher's exact test (Fisher, January 1922) on 2×2 tables comparing true positives and false positives across methods.

$$P = \frac{TP}{TP + FP} \tag{20}$$

Recall (R) (21) was estimated using a random sample of 300 interactions per site, rather than the full set of 32,024 (Location A) and 7,742 (Location B) detected interactions, balancing feasibility with statistical validity. To estimate the number of undetected conflicts (false negatives), the sample was analyzed using the Clopper–Pearson exact binomial interval (Clopper and Pearson, 1934), a method suited for rare-event evaluation.

$$R = \frac{TP}{TP + FN} \tag{21}$$

3.2.2. Benchmark comparison

For baseline comparison, a standard surrogate safety approach using 2D PET was implemented, applying a 1-second PET threshold to object center trajectories in BEV space (Peesapati et al., January 2013). As an additional baseline, PET was also computed in 3D using full object volumes, providing more accurate interaction assessment while maintaining the same threshold for comparability.

3.2.3. Module evaluation

Clustering, rule extraction and maneuver classification were evaluated independently. Cluster shapes and spatial patterns were compared to known traffic movements. All three modules were also validated using the same 300-sample subset as for conflict detection recall, with standard errors for accuracy calculated following Moore et al. (Moore et al., 2014). In addition, the influence of rule detection and evasive maneuver classification on the number of detected conflicts was analyzed, the LOF threshold was systematically varied, and the conflict precision obtained from LSTM- and transformer-based encoders was compared.

4. Results and Discussion

4.1. Qualitative analysis of reports

At both sites, Knowledge PET3D reliably detected rule violations and evasive maneuvers. The following examples highlight common conflict patterns and the benefits of integrating rule and behavior analysis.

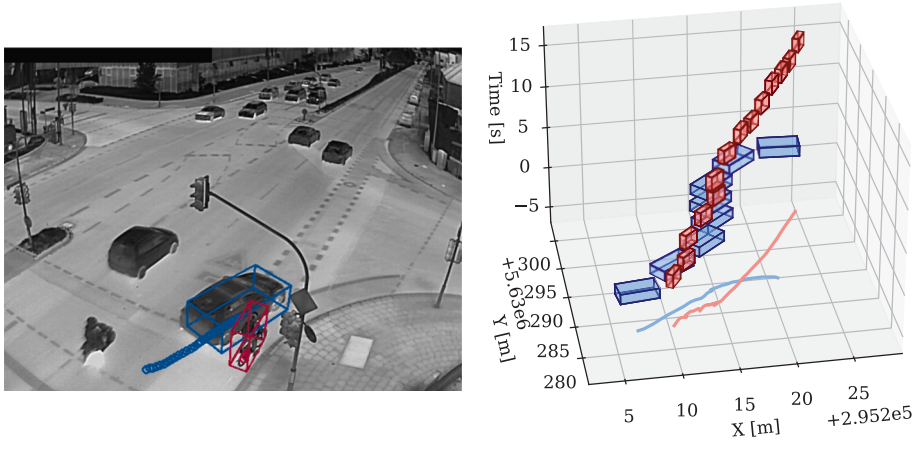
4.1.1. Qualitative analysis – Location A

4.1.1.1. Right-Turning Vehicle vs. Cyclist. A cyclist swerves and stops to avoid a right-turning vehicle (Fig. 8 and supplementary Video 6a). There is no immediate overlap between their paths. The intersection between the objects bottom areas, measured as an IoU greater than 0.1, occurs later with a PET of 2.07 s. This timing exceeds standard PET thresholds. As a result, PET2D and PET3D do not detect the conflict because the critical moment happens before the measured overlap. In contrast, Knowledge PET3D correctly identifies the traffic rule violation and detects the cyclist's evasive maneuver, indicated by a LOF score of 21. This confirms the presence of a high-risk interaction and demonstrates the importance of combining rule reasoning with behavioral indicators.

4.1.1.2. Left-Turning Vehicle vs. Oncoming Traffic. A left-turning vehicle cuts across an oncoming car (Fig. 9 and supplementary Video 7a), which slows to avoid collision. Although the PET is 1.7 s and the conflict appears minor, Knowledge PET3D detects rule violation and evasive behavior (LOF: 3.56), offering valuable insight for improving signal timing or intersection design.

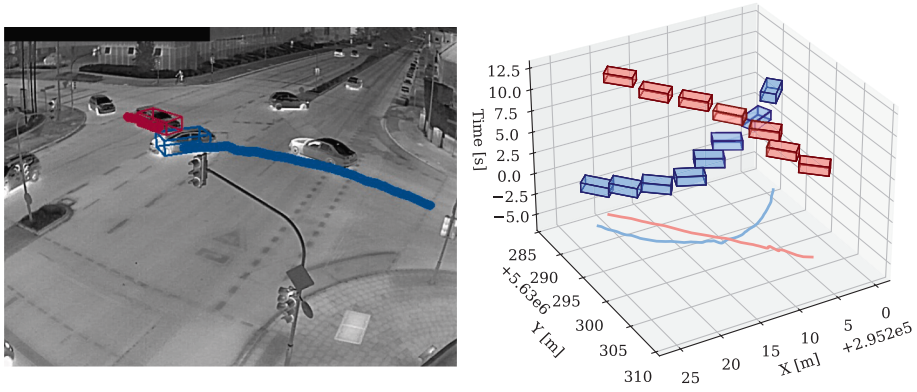
4.1.2. Qualitative analysis – Location B

4.1.2.1. Off-Ramp Conflict: Cyclist vs. Turning Vehicle. At Location B, several conflicts involve vulnerable road users (VRUs) yielding despite having the right-of-way (Fig. 10 and supplementary Video 8a). In this case, a bicycle swerves as a car passes without slowing. Although the PET is 2.13 s Knowledge PET3D flags the event due to rule violation and evasive behavior (LOF: 1,61). Such cases reveal ambiguous yielding behavior, likely from poor signage or visibility, and point to



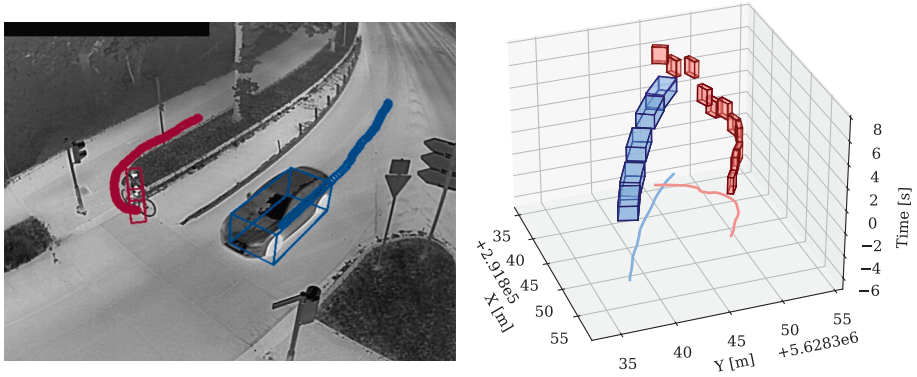
a) b)

Fig. 8. Most severe conflict: A right-turning car violates the bicycle's priority, narrowly avoiding collision through extreme evasive actions by both. (a) Thermal image shows the street-level context. The full video is available as supplementary material (Video 6a) (b) x–y trajectories illustrate spatial and temporal separation.



a) b)

Fig. 9. Frequent conflict at Location A: A left-turning car cuts off an oncoming vehicle, violating right of way. (a) Thermal image shows street context. The full video is available as supplementary material (Video 7a) (b) x–y trajectories reveal spatial and temporal gap.



a) b)

Fig. 10. Typical conflict at Location B: A car proceeds despite the bicycle's right of way, prompting the cyclist to swerve. Knowledge PET3D flags the evasive maneuver. (a) Thermal image shows street context. The full video is available as supplementary material (Video 8a) (b) x–y trajectories show spatial and temporal gap.

infrastructure improvement needs.

4.2. Quantitively comparison with existing methods

A central innovation of Knowledge PET3D is its maneuver-specific conflict classification, which leverages geometric trajectory analysis to distinguish between scenarios such as turning across paths, merging, and following. In contrast, PET2D and PET3D treat all trajectory overlaps uniformly. For a fair comparison, the same classification scheme was retrospectively applied to the PET2D and PET3D results.

4.2.1. Conflict overview

Table 1 summarizes conflicts by maneuver type. Knowledge PET3D detects fewer but more meaningful events, focusing on rule violations and evasive maneuvers. PET3D yields many proximity-only cases, especially in Following scenarios. Compared to PET2D, PET3D captures more overlaps by modeling full object volumes, often resulting in lower PET values. However, the volume of flagged interactions in PET3D limits interpretability, unlike the targeted output of Knowledge PET3D.

This filtering capability becomes more apparent when applying a 3-second threshold (Table 2). Even after excluding following and diverging events, PET2D and PET3D still produce thousands of cases (1444 for PET2D and 2829 for PET3D), making manual review

Table 1
Number of conflicts detected by different indicators across locations.

Conflict-Type	Loc A			Loc B		
	PET2D 1 s	PET3D 1 s	Knowledge PET3D	PET2D 1 s	PET3D 1 s	Knowledge PET3D
Following	49	1502	0	3	415	0
Turn-left across path	1	96	25	0	1	10
Turn-right across path	0	9	5	0	0	2
Crossing	6	23	6	0	0	0
Head-on	1	5	0	1	1	0
Merging	3	11	2	0	0	0
Diverging	1	64	0	0	0	0
All	61	1710	38	4	417	12

Table 2
Number of conflicts detected using PET2D and PET3D with a 3-second threshold, with and without including following and diverging conflicts.

Conflict-Type	Loc A			Loc B		
	PET2D 3 s	PET3D 3 s	Knowledge PET3D	PET2D 3 s	PET3D 3 s	Knowledge PET3D
All	9068	32,024	38	3111	7742	12
All w\ Following, Diverging	690	1735	38	754	1094	12

impractical and further underscoring the need for targeted filtering as implemented in Knowledge PET3D.

4.2.2. Crossing / turn-left across / turn-right across conflicts

Manual video review was conducted for conflicts involving crossing, turn-left-across-path, and turn-right-across-path scenarios. A conflict was confirmed if an interaction between road users posed a discernible safety risk. This included situations where a crash was plausibly avoided due to evasive actions, even when the road user had formal right-of-way. In contrast, detections were labeled as false positives if no observable conflict or behavioral adaptation occurred.

Table 3 reports the number of TPs and FPs detected at both locations using the three conflict detection methods. PET2D failed to detect any true positives due to its reliance on center-point logic, which generally results in higher PET values, and the strict 1-second threshold. PET3D improves detection by accounting for full object volumes, enabling earlier identification of overlaps. However, at Location A, this leads to a substantial increase in false positives. Many of these are from detection inaccuracies when vehicles pass in close proximity without actual interaction. At Location B, where such close passing events are rare, PET3D produced substantially fewer false positives.

At Location A, Knowledge PET3D achieved a significantly higher precision (22.22 %) than PET3D (2.34 %), confirmed by Fisher's exact test ($p = 0.00027$). At Location B, PET3D yielded perfect precision (100 %) with fewer detections, while Knowledge PET3D achieved 75 % precision. However, this difference was not statistically significant ($p = 0.769$), reflecting the small sample size and the absence of false positives in PET3D's detections. Knowledge PET3D consistently identifies a broader range of safety-relevant conflicts while keeping false positives at a manageable level. This is achieved through the integration of rule violation detection and behavior-based anomaly analysis, which enhances both interpretability and precision. The system filters out low-severity proximity events and focuses on interactions that pose a genuine safety risk.

Knowledge PET3D captures evasive actions, that increase absolute PET values and would be excluded by conventional threshold-based methods. These actions often indicate actual conflicts, even when the PET is high. Unlike PET2D and PET3D, which rely on rigid numerical cutoffs, Knowledge PET3D applies a knowledge-based filtering approach informed by traffic engineering principles. This leads to a higher true positive rate, offering greater value for safety analysis by capturing subtle or non-normative interactions.

However, the system is not without limitations. Variations in false positive rates between locations are partly due to detection inaccuracies, particularly in estimating object volumes, similar to PET3D. A second limitation arises from the use of a 3-second temporal window to define potential conflicts. While empirically chosen to balance recall and interpretability, this threshold can lead to cases where a rule violation is flagged despite no actual interaction. For instance, in one case at Location B (Fig. 11), a priority violation was detected between a bicycle and a car, although the car had already cleared the conflict zone. Although anomaly detection is used to confirm whether the rule violation affected the other road user, rare misclassifications remain when behavior changes are unrelated to interaction.

Nevertheless, the overall balance between detected true positives and manageable false positives demonstrates a favorable trade-off. Knowledge PET3D enables the detection of meaningful but often overlooked risks, supporting its use as a decision support tool in traffic safety assessments.

4.2.3. Following conflicts / head-on conflicts / Diverging conflicts

A key difference between PET2D/PET3D and Knowledge PET3D is the treatment of Following conflicts. PET3D flags many such cases (see Table 1), while Knowledge PET3D excludes them due to the absence of rule violations. Although close following can be risky, it is common in urban traffic and rarely indicates location-specific danger. As seen in Fig. 12 (Location A), these interactions are diffuse and overwhelm manual review. By filtering out low-severity, high-frequency events, Knowledge PET3D improves interpretability and focuses on actionable conflicts relevant for planning.

The small number of head-on conflicts detected by PET2D and PET3D were false positives, mostly caused by tracking inaccuracies. While following conflicts are not currently modeled, the modular framework allows for future extensions by for example adding a logic for

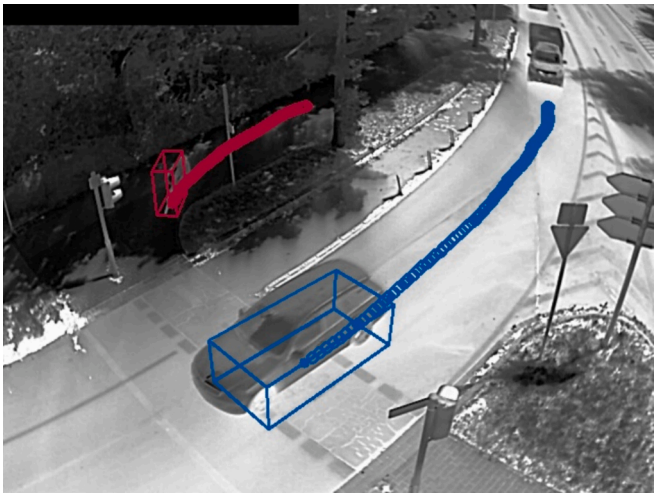


Fig. 11. Falsely detected priority violation: The car reaches the conflict point significantly earlier, so proceeding first does not constitute a rule violation.

Table 3
Number of true positive and false positive detections of crossing, turn-left across path or turn-right across path conflicts, evaluated at both locations. Results are shown for all indicators along with the resulting precision for each.

Conflict-Type	Loc A			Loc B		
	PET2D 1 s	PET3D 1 s	Knowledge PET3D	PET2D 1 s	PET3D 1 s	Knowledge PET3D
TP	0	3	8	0	1	9
FP	7	125	28	0	0	3
Precision [%]	0	2.34	22.22	/	100	75

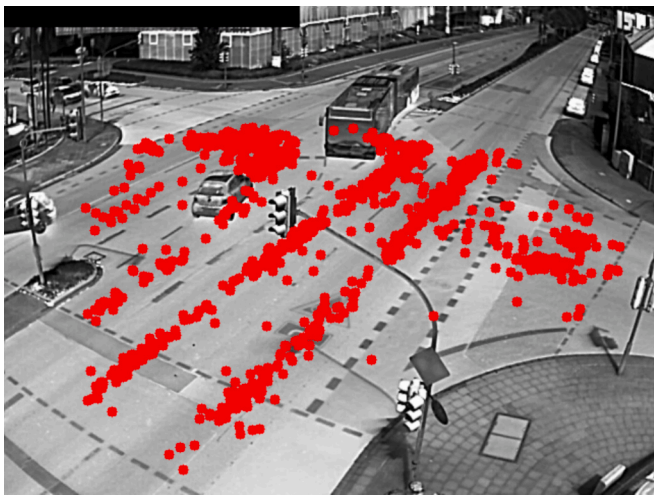


Fig. 12. Locations of following conflicts with PET3D < 1 s, projected onto the image to provide street context.

tailgating on highways.

4.2.4. Merging conflicts

Prior research has shown that PET is better suited for analyzing crossing conflicts than merging or rear-end scenarios, due to its reliance on sharply defined spatiotemporal proximity (Mahmud et al., December 2017). Therefore, merging conflicts were not a focus in the selected locations, and only a few such cases were observed. As shown in Table 4, no true positives were detected by any method. Knowledge PET3D produced only two false positives, compared to eleven from PET3D and three from PET2D. This suggests that Knowledge PET3D effectively suppresses non-critical merging events, avoiding overflagging.

However, due to the limited number of merging cases, no firm conclusions can be drawn about detection performance. Future studies should focus on locations with frequent merging, such as on-ramps, to further evaluate this aspect.

4.2.5. Estimating the number of missed conflicts

To assess detection accuracy, a sample of 300 interactions with PET3D ≤ 3 s was manually reviewed (Table 5). Since following conflicts dominate PET3D outputs (>70 %) but are excluded by design in Knowledge PET3D (Section 4.2.3), uniform sampling would bias the results. Instead, a stratified approach ensured balanced representation of relevant maneuver types, primarily crossing, turning, and merging, while limiting following and diverging conflicts to 10 %.

For recall estimation, only non-following and non-diverging conflicts were considered (resulting in 270 interactions per location). At Location A, 2 confirmed conflicts in the sample translate to an estimated total of 13 (90 % Confidence Interval (CI): 3–38). PET3D detected 3, resulting in an estimated recall of 23.1 % (90 % CI: 7.9–100 %), while Knowledge PET3D detected eight, yielding 61.5 % recall (90 % CI: 21.0–100 %) (Table 3). At Location B, one confirmed conflict corresponds to an estimated total of 4 (90 % CI: 0.3–15). PET3D detected one resulting in 25 % recall (90 % CI: 6.67 %-100 %), whereas Knowledge PET3D flagged nine true positives, exceeding the estimate and supporting a recall near 100 % (90 % CI: 60–100 %) (Table 3).

Table 5
Number of true positive and false positive detections of crossing, turn-left or turn-right across conflicts for all indicators at both locations.

Indicator	Loc A				Loc B			
	TP	FP	TN	FN	TP	FP	TN	FN
PET 3D	0	56	242	2	0	30	269	1
Knowledge PET3D	0	6	292	2	1	0	299	0

These findings confirm that Knowledge PET3D captures the majority of relevant conflicts, significantly outperforming PET3D in recall. While the small sample size limits precise estimates of absolute conflict numbers, the results provide a meaningful indication that few relevant events go undetected. More importantly, they highlight that PET3D misses a larger share of safety–critical interactions than Knowledge PET3D. The improved sensitivity reflects Knowledge PET3D's targeted design, which prioritizes interpretable events involving rule violations or evasive maneuvers.

4.3. Detailed study on key components

4.3.1. Influence Knowledge-PET3D components

Knowledge PET3D integrates rule interpretation and anomaly detection to filter interactions into meaningful conflicts. Table 6 summarizes the effect of disabling these components. The experiments were conducted on the already described 300 randomly sampled interactions per location.

To quantify the impact of these filters, we used Fisher's exact test to compare the change in the False Positive (FP) ratio against the baseline (full Knowledge PET3D). Disabling rule interpretation resulted in no additional true positives but a highly statistically significant increase in false positives at Location A (37 vs. 6, $p \approx 7x10^{-7}$). At Location B, the increase was not statistically significant (2 vs. 0, $p \approx 0.499$). This suggests that at the simpler site (Location B), the behavioral consequence (i.e., the evasive action) is the dominant signal, which is already robustly captured by the anomaly detection module. This makes the explicit rule filter less critical in this specific context.

Excluding anomaly detection yielded one additional true positive at Location A but led to a highly statistically significant rise in false positives at both locations (Location A: 25 vs. 6, $p \approx 0.00064$; Location B: 15 vs. 0, $p \approx 5.1x10^{-5}$). These results quantitatively confirm that both modules act primarily as essential filters: rule interpretation suppresses interactions without priority violations, while anomaly detection excludes those lacking behavioral deviations. Their combination is statistically necessary to achieve the high precision required for engineering practice.

Table 6
Impact of rule interpretation and evasive action detection on conflict detection results for 300 random samples at both locations.

Indicator	Loc A				Loc B			
	TP	FP	TN	FN	TP	FP	TN	FN
Knowledge PET3D w\ evasive action	1	25	273	1	1	15	284	0
Knowledge PET3D w\ rule	0	37	261	2	1	2	297	0
Knowledge PET3D	0	6	292	2	1	0	299	0

Table 4
Number of true positive and false positive detections of merging conflicts for all indicators at both locations, along with the resulting precision.

Conflict-Type	Loc A			Loc B		
	PET2D 1 s	PET3D 1 s	Knowledge PET3D	PET2D 1 s	PET3D 1 s	Knowledge PET3D
TP	0	0	0	0	0	0
FP	3	11	2	0	0	0
Precision	0 %	0 %	0 %	/	/	/

4.3.2. Clustering

Fig. 13 and Fig. 14 visualize the resulting trajectory clusters for Locations A and B, respectively, with each cluster representing a distinct movement pattern in one direction. Since the clustering was performed directionally, each movement pattern appears twice, once for each direction of travel.

Overall, the clustering algorithm successfully identified all major movement patterns, including those at the more complex intersection in Location A. At Location A, some turning maneuvers are occasionally grouped into one cluster due to vehicles using similar center lanes when turning. In Location B, some short crossing trajectories are absorbed into nearby clusters (e.g., Cluster 2), but this has no significant impact. Clustering is mainly used to assign maneuver types, interpret traffic rules, and detect anomalies, where functional similarity is more important than precise geometric separation.

Table 7 summarizes cluster assignment accuracy based on 600 manually reviewed samples from interactions with PET3D < 3 s. The framework achieves 87.98 % accuracy at Location A and 98.17 % at Location B, totaling 92.92 %. Standard errors and confidence intervals were computed following Moore et al. (2014) (Moore et al., 2014). The narrow 90 % confidence intervals indicate statistical validity of evaluation on the 600 samples subset.

4.3.3. Maneuver classification

As shown in Table 8, Maneuver classification, which is based on assigned clusters, achieved 77.7 % accuracy at Location A and 96.3 % at Location B, resulting in 87.0 % overall. The lower accuracy at Location A reflects geometric complexity and ambiguous endpoints. The tight confidence intervals further support the statistical reliability of these results.

4.3.4. Rule interpretation

As shown in Table 9, rule interpretation accuracy reached 96.5 % overall, with only minor misclassifications caused by contextual ambiguities. The high accuracy is primarily due to the larger 3-second time window, which increases the robustness of priority assessments. Narrow confidence intervals at both sites underscore the method's consistency and reliability in varied urban settings.

4.3.5. Transformer based evasive action detection

The anomaly detection model was trained for 100 epochs with a batch size of 16, using the AdamW optimizer (learning rate = 1e-4). The learning rate was reduced by a factor of 0.5 whenever the validation loss plateaued for 5 consecutive epochs. Each input trajectory was capped at 1200 steps and augmented by randomly masking approximately 40 % of displacement vectors in contiguous segments (up to 5 steps) and injecting Gaussian noise ($\sigma = 0.1$) into the remainder. The best model was selected based on minimum validation loss, which was reached at epoch 56. Fig. 15 shows stable convergence throughout training. Example reconstructions in Fig. 16 demonstrate accurate modeling of typical motion patterns, with minor limitations in capturing trajectories involving curves.

4.3.6. Comparison of transformer and LSTM encoders

Both encoders achieved nearly identical trajectory reconstruction performance (RMSE 2.7 cm for the transformer and RMSE 2.6 cm for the LSTM) under the given preprocessing conditions. To rigorously evaluate the final conflict detection performance, we applied Fisher's exact test to the True Positive (TP) and False Positive (FP) counts shown in Table 10. At Location A, the performance differences were marginal (22.86 % vs 21.05 % precision), and this difference is not statistically significant ($p \approx 1.0$).

In contrast, at Location B, the performance difference is highly statistically significant ($p \approx 0.016$). The LSTM yielded slightly more true positives (11 vs. 9) but also a considerably and statistically significant higher number of false positives (24 vs 3). This finding is practically meaningful for engineering applications. The primary goal of the framework is to deliver a precise, verifiable set of conflicts. The LSTM's high false positive rate (24 FPs) would unacceptably increase the human review burden, making the Transformer's 75.00 % precision vastly superior in practice. Location B contains a diverse cluster of left-turning vehicles (cluster 3) with similar geometric paths but varying temporal

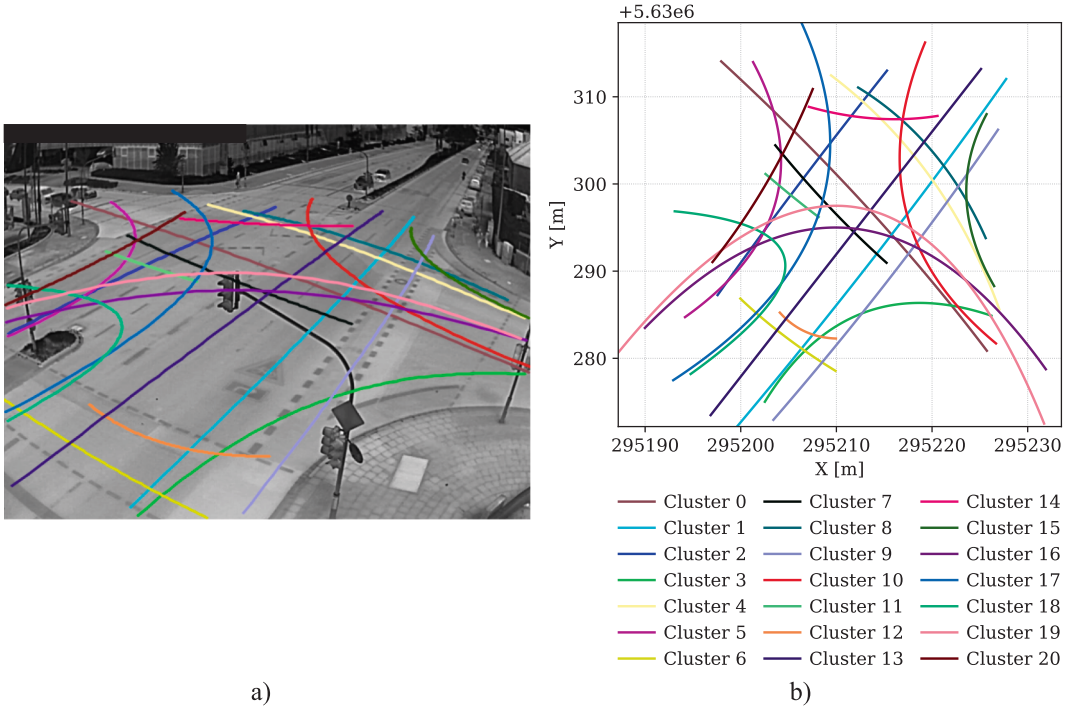


Fig. 13. Mean cluster trajectories Location A: (a) Thermal image in pixel coordinates showing the street context; (b) Visualization of the same cluster means in UTM coordinates, illustrating the distinction between clusters.

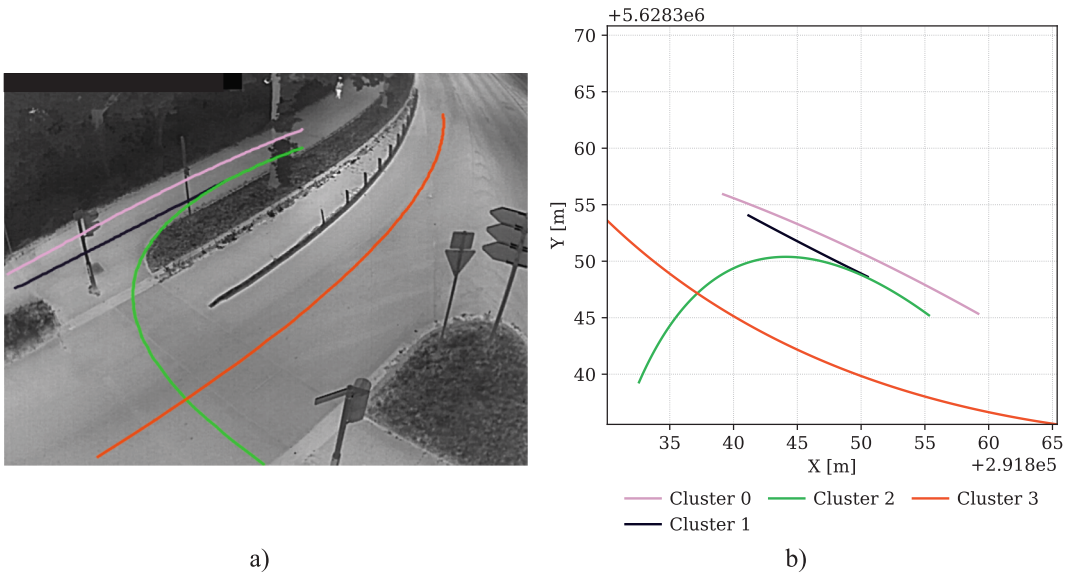


Fig. 14. Mean cluster trajectories Location B: (a) Thermal image in pixel coordinates showing the street context; (b) Visualization of the same cluster means in UTM coordinates, illustrating the distinction between clusters.

Table 7
Number of true positive and false positive detections of crossing, turn-left or turn-right across conflicts for all indicators at both locations, along with the resulting precision.

Location	# Correct assigned clusters (TP)	# False assigned clusters (FP)	Accuracy	90 % CI (+- SE)
Loc A	526	74	87.98 %	85.4 % – 90.5 %
Loc B	589	11	98.17 %	97.1 – 99.1
Total	1115	85	92.92 %	91.5–94.3

Table 8
Evaluation of maneuver assignment, showing correctly and incorrectly assigned clusters, along with the resulting precision.

Location	# Correct maneuvers	# False maneuvers	Precision	90 % CI (+- SE)
Loc A	233	67	77.67 %	73.4–82.0
Loc B	289	11	96.33 %	94.4–98.2
Total	522	78	87.00 %	84.3–89.7

Table 9
Number of true positive and false positive detections of crossing, turn-left or turn-right across conflicts for all indicators at both locations, along with the resulting precision.

Location	# Correct priority	# False priority	Precision	90 % CI (+- SE)
Loc A	289	11	96.33 %	94.4–98.2
Loc B	290	10	96.67 %	94.9–98.4
Total	579	21	96.5 %	95.1–97.9

and kinematic characteristics. The Transformer's statistically superior performance suggests it can better handle such heterogeneity by implicitly capturing broader contextual dependencies. Overall, this analysis provides a clear, statistically significant justification for the Transformer's advantage in complex scenarios, validating its use in the proposed framework. This confirms the Transformer is the more robust and reliable architecture for the framework, as it performs comparably in simple cases and significantly better in complex ones. Further studies across additional sites and with tailored training strategies could help validate these interpretations.

4.3.7. *Lof-based anomaly detection*

Fig. 17 exemplary visualizes anomaly classification in cluster 2 (bicycles) at Location B, including the critical conflict highlighted in Section 4.1.2. These results illustrate the model's capability to detect behavioral deviations such as abrupt turns or stop-and-go movements, providing a valuable complementary filter for identifying ambiguous interactions.

In the current implementation, LOF is applied to bins of trajectories grouped by cluster and object class across the full observation period (approximately one week). This design ensures that sufficient samples are available for local density estimation, but it also implies sensitivity to environmental factors such as rain, fog, or icy surfaces or changes in traffic infrastructure, which may alter the distribution of normal behavior. The stability of LOF estimates is known to scale with the number of available trajectories (Breunig et al., 2000); since larger samples provide more reliable local density neighborhoods.

To examine threshold sensitivity, Table 11 reports results for the above mentioned 300 randomly sampled conflicts per location, while Table 12 summarizes all detected conflicts. Lower thresholds (e.g., LOF = 1.2) increase the number of detected conflicts (eg. 12 vs 9 at location B), but generate more false positives (eg. 47 vs 30 on Location A), whereas higher thresholds (e.g., LOF = 1.8) reduce false positives (e.g., from 30 to 20 at Location A, and from 3 to 1 at Location B) while slightly decreasing true positives (9 vs 6 at Location B). Precision improves accordingly (up to 86 % at Location B for LOF = 1.8), but at the cost of sensitivity.

To test if these trade-offs are statistically robust, we applied Fisher's exact test to the (TP, FP) counts in Table 12. At Location B, the large apparent precision jump from 50 % (LOF 1.2) to 75 % (LOF 1.5) is not statistically significant ($p \approx 0.282$) nor is the jump from 75 % (LOF 1.5) to 86 % (LOF 1.8) ($p \approx 1.0$). Similarly, the differences at Location A were also not statistically significant ($p > 0.56$). This quantitatively confirms that the default threshold of 1.5, which was not optimized on the test data to avoid introducing dataset-specific bias, is a robust and stable choice. The observed precision changes across thresholds are not statistically significant, validating our approach of preserving independence between evaluation data and hyperparameter choice. Future work should expand the data basis beyond one week and consider context-specific bins (e.g., weather conditions, time of day) to further improve

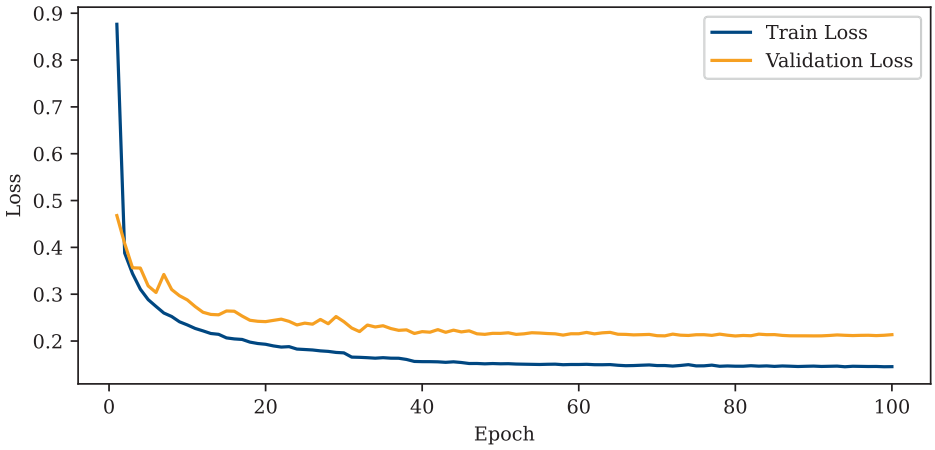


Fig. 15. Training and validation loss of the transformer network. A sharp decrease is observed during the first epoch, followed by convergence.

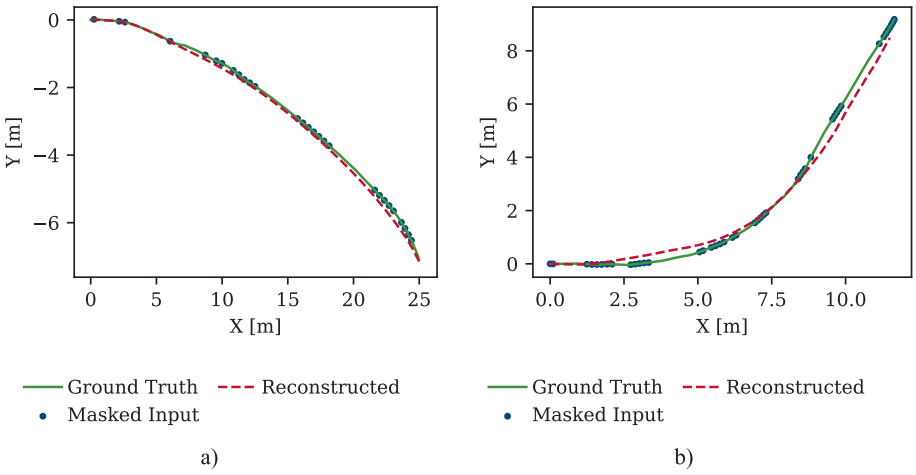


Fig. 16. Example of reconstructed trajectories from the validation set: (a) shows a well-fitting reconstruction, while (b) illustrates minor inaccuracies in the curve.

Table 10
Comparison of true and false positive conflict detections between the LSTM and transformer-based encoders at both locations.

Conflict Type	Loc A LSTM based	Transformer based	Loc B LSTM based	Transformer based
TP	8	8	11	9
FP	27	30	24	3
Precision	22.86 %	21.05 %	31.43 %	75.00 %

robustness and generalizability.

4.4. Practical usability and exemplary presentation of the engineering findings

4.4.1. Location A

Location A is a complex urban intersection with high multimodal traffic. Knowledge PET3D highlights interpretable, rule-relevant conflicts that offer practical value for engineering assessment.

A frequent conflict involved left-turning vehicles misjudging gaps in oncoming traffic. Video analysis confirmed that these situations were not due to visibility limitations but rather to driver impatience or poor judgment. Adjusting signal timing to separate turning and through movements could help reduce these risks.

Conflicts between right-turning vehicles and cyclists were also common. In these cases, video review showed that limited driver

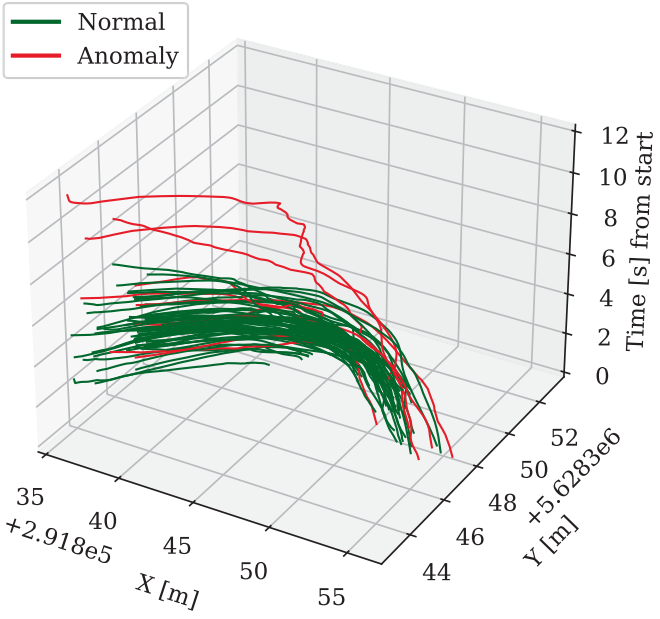


Fig. 17. Trajectories of bicycles from Cluster 3 at Location B, which have interactions with vehicles from Cluster 2 without traffic rule violation. Detected anomalies are highlighted in red.

Table 11
Number of true positive, false positive, true positive and true negative detections for different LOF values for 300 random samples at both locations.

LOF Value	Loc A				Loc B			
	TP	FP	TN	FN	TP	FP	TN	FN
LOF 1.2	0	10	288	2	1	2	297	0
LOF 1.8	0	3	295	2	0	0	299	1
Baseline LOF 1.5	0	6	292	2	1	0	299	0

Table 12
Number of true positive and false positive detections for all detected conflicts with differen LOF values at both locations, along with the resulting precision.

LOF Value	Loc A			Loc B		
	1.2	Baseline1.5	1.8	1.2	Baseline 1.5	1.8
TP	9	8	8	12	9	6
FP	47	30	20	12	3	1
Precision	19 %	21 %	40 %	50 %	75 %	86 %

awareness, potentially caused by cyclist positioning and lack of attention, was a key factor. Engineering measures such as improved road markings, clearer signage, or relocating bike lanes closer to pedestrian crossings could enhance visibility and reduce conflict likelihood.

4.4.2. Location B

At Location B, Knowledge PET3D revealed multiple instances where drivers failed to yield to cyclists, despite clear visibility. While PET metrics alone might suggest visibility problems, video analysis showed that the true issue was a lack of rule compliance. Drivers often disregarded cyclists' right of way, likely due to unclear priority signaling at the intersection.

To address this, interventions such as reinforced signage or pavement markings could clarify priority and help reduce future conflicts.

4.5. Computational performance

To assess the feasibility of long-term monitoring, the computational performance of all major framework components was benchmarked. Experiments were conducted on a workstation with an NVIDIA Quadro RTX 5000 GPU and an Intel Xeon E5-2640 v3 CPU (2.60 GHz, 16 cores).

The video processing layer, responsible for detection, tracking and trajectory enhancement, constitutes the primary computational bottleneck. It processed approximately 24 h of video (30 fps) from Location A in 49,305 s (49.2 fps on average) and from Location B in 46,311 s (52.4 fps on average), achieving processing speeds more than 1.6 times faster than real time. All subsequent analysis modules are computationally lightweight. Their processing times for 24 h of data (or one-time setup operations) are summarized in Table 13. The higher throughput at Location B is attributed to lower traffic density and fewer detected object clusters.

The results confirm the framework's operational feasibility. The main processing layer exceeds real-time performance, while subsequent modules remain highly efficient. Further optimization through parallelization, such as multi-threaded processing of multiple video streams, is expected to enhance throughput for large-scale deployments.

Table 13
Computational time for analysis modules.

Module	Location A [s]	Location B [s]	Data scope
Trajectory clustering	141	23	One-time (24 data)
Cluster assignment	450	45	Per 24 h data
Interaction detection	4487	2104	Per 24 data
Rule extraction	0.3	0.3	One-time (24 data)
Rule assignment	77	20	Per 24 h data
LOF Calculation	1327	33	One-time (1 week data)

4.6. Discussion on Generalizability and Limitations

4.6.1. Generalization and transferability

The evaluation covered two distinct intersections, one complex multilane signalized junction and one simpler yield-controlled slip lane, representing typical inner-city conditions. The latter, located on the city's outskirts, approaches a semi-rural configuration, demonstrating robustness to varied geometry and regulation. Since all components after object detection operate in georeferenced world coordinates, the framework is largely independent of specific sensors or camera setups and can be adapted to new sites without retraining the full pipeline. The framework's key mechanism for practical transferability is the self-supervised transformer. This model is intentionally designed to be rapidly adapted to a new site by training only on the first 24 h of unlabeled data from that location. This self-learning approach is a practical alternative to full cross-location generalization, as it allows the anomaly detection module to learn the local, nominal motion patterns of a new intersection without requiring labeled data or a pre-trained universal model. While this paper demonstrates the framework's robustness at two distinct sites, a full quantitative cross-location evaluation to assess transferability and explore self-learning domain adaptation remains a high-priority direction for future work.

The study focuses on identifying the causes of traffic conflicts, which in inner-city settings often arise from infrastructure or traffic management issues. Following conflicts and therefore most highway scenarios are excluded by design, as these longitudinal interactions are better represented through other surrogate safety measures such as Time to Collision (TTC) (Singh et al., March 2024).These events typically reflect individual driver risk-taking rather than systemic design faults. While excluding them may introduce bias in broader safety assessments, the framework's goal is not to compare overall safety levels but to uncover the root causes of critical urban conflicts.

4.6.2. Conflict scope and validation

Conflict validation was conducted by a single domain expert to ensure consistent interpretation across the dataset. Although this may introduce limited subjectivity, the framework relies on objective indicators including spatiotemporal interaction, rule violation, and detected evasive action, which reduce personal bias and ensure reproducibility. Unlike many prior works that rely on crash data or purely proximity-based metrics, this approach provides interpretable behavioral reasoning. Future studies may extend this with multi-rater validation or long-term crash correlation, though such efforts demand extensive observation and raise ethical and logistical constraints.

5.3. Behavioral and normative rule interpretation

A critical aspect of the framework is its handling of traffic rules, which Section 2.5.2 describes as operating in two distinct modes. The framework's default, automatic mode infers the de facto (behavioral) rule using Eq. (19). This mode is intended for rapid, interpretable analysis and performs well in environments with general compliance (96.5 % accuracy). However, as the framework requires a 'Rule Violation' to detect a conflict, this default mode is vulnerable to bias in environments with frequent non-compliance. It would incorrectly learn the unsafe norm, causing the 'Rule Violation' check to fail.

This is why the methodology includes the 'Fixed Prior' mode as the intended, robust safeguard for engineering safety assessment. In this operational mode, the engineer supplies the known de jure (legal) rule as a fixed ground truth. The system bypasses behavioral inference Eq. (19) entirely and uses this legal prior for its violation check Eq. (18). This intended operational mode is immune to the bias from unsafe behavioral norms and robustly detects all legal violations. The de facto estimate Eq. (19) can then be used as a valuable diagnostic to quantify

the (unsafe) discrepancy between behavior and law.

5.4. Broader methodological transferability

Beyond the field of traffic safety, the presented framework exemplifies the broader concept of knowledge-informed AI, where domain expertise is systematically integrated into data-driven models to enhance interpretability and robustness under data scarcity and heterogeneity. Similar principles have been explored in other engineering domains, such as energy systems modeling (Tao et al., 2023), highlighting the growing relevance of hybrid approaches that combine learning and reasoning. This connection underlines the transferability of our framework's methodological principles to other data-constrained applications.

6. Conclusion

This work presents Knowledge PET3D, a novel framework for interpretable near-miss detection in thermal traffic video, bridging the gap between raw sensing data and actionable engineering insights. The key contributions and outcomes of this study are as follows:

- A three-layered framework integrating 3D thermal trajectory extraction, rule-aware interaction filtering, transformer-based anomaly detection, and an engineering interpretation layer that provides structured, video-backed conflict reports for domain experts, enabling privacy-compliant, scalable, and interpretable traffic safety assessments.
- A knowledge-based extension of Post-Encroachment Time (PET), which incorporates traffic rules and behavioral cues to filter conflicts and improve interpretability and engineering relevance of detected near-miss events.
- A clustering and maneuver classification method using downsampled trajectory geometry and angular analysis, achieving about 93 % clustering accuracy (90 % CI: 91.5 %-94.3 %) and 87.0 % (90 % CI: 84.3 %-89.7 %) maneuver classification accuracy, enabling semantic understanding of traffic interactions.
- A transformer-based self-supervised anomaly model that learns behavioral deviations without labeled evasive actions, enabling the detection of subtle evasive maneuvers such as swerving or abrupt braking.
- A rule interpretation module that maps cluster-based motion patterns to priority logic, achieving over 96.5 % (90 % CI: 95.1 %-97.9 %) accuracy in identifying rule violations across two urban locations.

Compared to baseline methods using conventional 2D and 3D PET thresholds, Knowledge PET3D reduces false positives by up to 93 % and detects over 400 % more true positives at complex intersections, capturing critical interactions often missed by rigid spatiotemporal approaches.

Although not yet optimized for real-time edge deployment, the system is well-suited for offline analysis. Future work will focus on real-time deployment and scalability across diverse traffic scenes. This includes support for additional conflict types such as lane changes and rear-end collisions, integration of multi-sensor inputs to expand field of view, and validation against real crash data. Future work will also explore cross-location generalization, assessing how models trained at one site perform when applied to another. Improvements in trajectory modeling and continuous learning across locations will further strengthen its role as a decision support tool for proactive road safety assessment.

CRediT authorship contribution statement

Arnd Pettirsch: Methodology, Software, Data curation, Formal analysis, Investigation, Funding acquisition, Writing – original draft. **Alvaro Garcia-Hernandez:** Supervision, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The study was part of the research project 'SmarteAmpel' founded by the German Federal Ministry of Economic Affairs and Climate Action (BMWK) (grant no KK5292303ER1). Data was further collected in the project 'AdaptIn' founded by the German Federal Ministry of Education and Research (BMBF) (grant no 16SV8670).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2025.130821.

Data availability

The code and dataset supporting this study's findings are publicly available. The analysis code is hosted in a GitHub repository at: https://github.com/4rnd25/knowledge_PET3D. The comprehensive dataset, which includes 10 example thermal videos, all processed trajectory data, final conflict data, and intermediate results (e.g., clusters, priorities), is permanently archived on Mendeley Data (https://doi.org/10.17632/2wtbb9jnpx.1).

References

Abdel-Aty, M., Wu, Y., Zheng, O., & Yuan, J. (2022). Using closed-circuit television cameras to analyze traffic safety at intersections based on vehicle key points detection. *Accident Analysis & Prevention, 176*, Article 106794. https://doi.org/10.1016/j.aap.2022.106794

Federal Highway Administration (FHWA), 'Road Safety Audit Guidelines', Washington, 2006. https://highways.dot.gov/safety/data-analysis-tools/rsa/fhwa-road-safety-audit-guidelines.

T. Alldieck, C. Bahnsen and T. B. Moeslund, 'Context-Aware Fusion of RGB and Thermal Imagery for Traffic Monitoring', Sensors, Vol. 16, Article 1947, November 2016. https://doi.org/10.3390/s16111947.

B. L. Allen, B. T. Shin and P. J. Cooper, 'Analysis of Traffic Conflicts and Collisions', Analysis of traffic conflicts and collisions (No. HS-025 846), Washington, D.C., 1978.

A. S. Bhadoriya, V. Vegamoor and S. Rathinam, 'Vehicle Detection and Tracking Using Thermal Cameras in Adverse Visibility Conditions', Sensors, Vol. 22, Article 4567, 2022. https://doi.org/10.3390/s22124567.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *in Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93–104). https://doi.org/10.1145/342009.335388

CAREC Road Safety Engineering Manual 1: Road Safety Audit, 2018. https://www.adb.org/publications/carec-road-safety-audit-engineering-manual.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*, 404–413. https://doi.org/10.1093/biomet/26.4.404

Axis Communications. (2021). AXIS Q1952-E Thermal Camera [Datasheet]. Axis Communications AB. Retrieved June 26, 2025, from https://www.axis.com/dam/public/b0/8b/93/datasheet-axis-q1952-e-thermal-camera-en-US-350847.pdf.

Directive 2008/96/EC of the European Parliament and of the Council of 19 November 2008 on Road Infrastructure Safety Management, 2008. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32008L0096.

Esri. (2025). ArcGIS web application – View application. Retrieved April 24, 2025, from https://www.arcgis.com/apps/View/index.html?appid=df7cee38677f479c8697026ebf920431.

M. Ester, H.-P. Kriegel, J. Sander, X. Xu and others, 'A density-based algorithm for discovering clusters in large spatial databases with noise', in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96), pp. 226-231, 1996.

Fisher, R. A. (January 1922). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society, 85*, 87. https://doi.org/10.2307/2340521

Hou, Q., Yang, Y., Liang, J., Huo, X., & Leng, J. (May 2025). A deep transfer learning approach for Real-Time traffic conflict prediction with trajectory data. *Accident*

Analysis & Prevention, 214, Article 107966. https://doi.org/10.1016/j.aap.2025.107966

Hydén, C. (1987). *The development of a method for traffic safety evaluation: The Swedish Traffic Conflicts Technique*. Department of Traffic Planning and Engineering: Bulletin Lund Institute of Technology.

Islam, Z., Abdel-Aty, M., Goswamy, A., Abdelraouf, A., & Zheng, O. (June 2023). Effect of signal timing on vehicles' near misses at intersections. *Scientific Reports, 13*. https://doi.org/10.1038/s41598-023-36106-3

C. Johnsson and A. Laureshyn, 'Identification of evasive manoeuvres in traffic interactions and conflicts', Traffic Safety Research, Vol. 3, 2022. https://doi.org/10.55329/erqd8683.

Kalman, R. E. (1960). 'A New Approach to Linear Filtering and Prediction Problems', Transactions of the ASME–Journal of Basic. *Engineering, 82*, 35–45. https://doi.org/10.1115/1.3662552

P. Kar, S. Kumar, S. Samalla, M. Chunchu and K. V. R. Ravi Shankar, 'Exploratory analysis of evasion actions of powered two-wheeler conflicts at unsignalized intersection', Accident Analysis & Prevention, Vol. 194, Article 107363, January 2024. https://doi.org/10.1016/j.aap.2023.107363.

A. Kim, A. Osep and L. Leal-Taixe, 'EagerMOT: 3D Multi-Object Tracking via Sensor Fusion', in Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 11315-11321, IEEE, 2021. https://doi.org/10.1109/ICRA48506.2021.9562072..

Laureshyn, A., Svensson, Å., & Hydén, C. (2010). Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. *Accident Analysis & Prevention, 42*, 1637–1646. https://doi.org/10.1016/j.aap.2010.03.021

Mahmud, S. M. S., Ferreira, L., Hoque, M. S., & Tavassoli, A. (December 2017). Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. *IATSS Research, 41*, 153–163. https://doi.org/10.1016/j.iatssr.2017.02.001

Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (March 2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research, 25*, Article 100113. https://doi.org/10.1016/j.amar.2020.100113

Moore, D. S., McCabe, G. P., & Craig, B. A. (2014). *Introduction to the practice of statistics* (8. ed., student ed.,). New York: Freeman.

Mukherjee, D. (2025). 'Assessing pedestrian safety at urban signalized intersections across various land use types: Insights from a mid-sized Indian city', Discover. *Applied Sciences, 7*, April. https://doi.org/10.1007/s42452-025-06945-y

Othman, I., & Ros, C. (2013). *A comparative review of road safety audit guidelines of selected countries*. UTM: Universiti Teknologi Malaysia. https://doi.org/10.11113/jt.v65.2148.

Peesapati, L. N., Hunter, M. P., & Rodgers, M. O. (January 2013). Evaluation of Postencroachment Time as Surrogate for Opposing Left-Turn Crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2386*, 42–51. https://doi.org/10.3141/2386-06

Pettirsch, A., & Garcia-Hernandez, A. (April 2025). New generation thermal traffic sensor: A novel dataset and monocular 3D thermal vision framework. *Knowledge-Based Systems, 315*, Article 113334. https://doi.org/10.1016/j.knosys.2025.113334

Pettirsch, A., & Garcia-Hernandez, A. (April 2025). Overcoming Data Scarcity in Roadside Thermal Imagery: A New Dataset and Weakly Supervised Incremental Learning Framework. *Sensors, 25*, 2340. https://doi.org/10.3390/s25072340

Rezaei, M., Azarmi, M., & Mir, F. M. P. (October 2023). 3D-Net: Monocular 3D object recognition for traffic monitoring. *Expert Systems with Applications, 227*, Article 120253. https://doi.org/10.1016/j.eswa.2023.120253

D. Singh and P. Das, 'A Review on Surrogate Safety Measures in Safety Evaluation and Analysis', in Proceedings of the Sixth International Conference of Transportation Research Group of India (CTRG 2021), Vol. 273, L. Devi, M. Errampalli, A. Maji and G. Ramadurai, (Eds.), Springer, Singapore, 2023. https://doi.org/10.1007/978-981-19-4204-4_7.

Singh, D., Das, P., & Ghosh, I. (March 2024). Conflict-Based safety evaluations at unsignalized intersections using surrogate safety measures. *Heliyon, 10*, Article e27665. https://doi.org/10.1016/j.heliyon.2024.e27665

European Data Protection Supervisor. (2023). Video surveillance [Guidelines]. European Data Protection Supervisor. Retrieved June 26, 2025, from https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/video-surveillance_en.

S. Tao, C. Sun, S. Fu, Y. Wang, R. Ma, Z. Han, Y. Sun, Y. Li, G. Wei, X. Zhang, G. Zhou und H. Sun, 'Battery Cross-Operation-Condition Lifetime Prediction via Interpretable Feature Engineering Assisted Adaptive Machine Learning', ACS Energy Letters, Vol. 8, p. 3269–3279, July 2023.https://doi.org/10.1021/acsenergylett.3c01012.

Tarko, A., Davis, G., Saunier, N., Sayed, T., & Washington, S. (2009). Surrogate Measures of Safety. In D. Lord, & S. Washington (Eds.), *Safe Mobility: Challenges, Methodology and Solutions (Transport and Sustainability* (Vol. 11, pp. 383–405). Leeds: Emerald Publishing Limited. https://doi.org/10.1108/S2044-994120180000011019.

P. P. Tasgaonkar, R. D. Garg and P. K. Garg, 'Vehicle Detection and Traffic Estimation with Sensors Technologies for Intelligent Transportation Systems', Sensing and Imaging, Vol. 21, Article 29, 2020. https://doi.org/10.1007/s11220-020-00295-2.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. u. Kaiser and I. Polosukhin, 'Attention is All you Need', in Advances in Neural Information Processing Systems, Vol. 30, 2017.

Wang, X., Shi, R., Leich, A., Saul, H., Sohr, A., & Bei, X. (2025). Conflict Extraction and Characteristics Analysis at Signalized Intersections Using Trajectory Data. *International Journal of Transportation Science and Technology*. https://doi.org/10.1016/j.ijtst.2024.12.002

Wei, Y., Li, K., & Tang, K. (February 2019). Trajectory-based identification of critical instantaneous decision events at mixed-flow signalized intersections. *Accident Analysis & Prevention, 123*, 324–335. https://doi.org/10.1016/j.aap.2018.11.019

X. Weng, J. Wang, D. Held and K. Kitani, '3D Multi-Object Tracking: A Baseline and New Evaluation Metrics', in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020. https://doi.org/10.1109/IROS45743.2020.9341164.